Operationally constrained quickest change detection with multiple post-change distributions

Lau, Tze Siong

2020

Lau, T. S. (2020). Operationally constrained quickest change detection with multiple post-change distributions. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/143188>

<https://doi.org/10.32657/10356/143188>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 19:57:02 SGT

Operationally Constrained Quickest Change Detection with Multiple Post-Change Distributions

Lau Tze Siong

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

15-01-2020

.....

Date

.....

Lau Tze Siong

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

16-01-2020

.....

Date

.....

Assoc. Prof. Tay Wee Peng

Authorship Attribution Statement

This thesis contains material from 2 papers published, 1 paper submitted in the following peer-reviewed journals and from 1 paper submitted to a conference in which I am listed as an author.

Chapter 3 is published as **T. S. Lau**, W. P. Tay, and V. V. Veeravalli, “A binning approach to quickest change detection with unknown post-change distribution,” *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 609 – 621, Feb. 2019.

The contributions of the co-authors are as follows:

- I proposed the idea, designed the algorithms, provided the theoretical results and prepared the manuscript drafts.
- Prof. Tay Wee Peng discussed the problem formulation and theoretical results with me.
- Prof. Tay Wee Peng and Prof V. V. Veeravalli revised the manuscript.

Chapter 4 is published as **T. S. Lau** and W. P. Tay, “Quickest change detection in the presence of a nuisance change,” *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5281 – 5296, Oct. 2019.

The contributions of the co-authors are as follows:

- I proposed the idea, designed the algorithms, provided the theoretical results and prepared the manuscript drafts.
- Prof. Tay Wee Peng discussed the problem formulation and theoretical results with me.
- Prof. Tay Wee Peng revised the manuscript.

Chapter 5 is submitted as **T. S. Lau** and W. P. Tay, “Asymptotically Optimal Sampling Policy for Quickest Change Detection with Observation-Switching Cost,” to the *IEEE Trans. Signal Process.*

The contributions of the co-authors are as follows:

- I proposed the idea, designed the algorithms, provided the theoretical results and prepared the manuscript drafts.
- Prof. Tay Wee Peng discussed the problem formulation and theoretical results with me.
- Prof. Tay Wee Peng revised the manuscript.

Chapter 6 is submitted as **T. S. Lau** and W. P. Tay, “Privacy-Aware Quickest Change Detection,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Barcelona, Spain, May 2020*.

The contributions of the co-authors are as follows:

- I designed the algorithms, provided the theoretical results and prepared the manuscript drafts.
- Prof. Tay Wee Peng proposed the idea, discussed the problem formulation and theoretical results with me.
- Prof. Tay Wee Peng revised the manuscript.

15-01-2020

.....

Date

.....

Lau Tze Siong

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Tay Wee Peng for accepting me into his group. During this period, he provided a rewarding graduate school experience by giving me intellectual freedom for my work, supporting me for conferences, providing fresh perspectives and encouragement when I meet difficult problems and demanding a high quality of work. Without his feedback and encouragement, this PhD would not have been achievable. I have benefited greatly from the regular group study session he organized despite his busy schedule. His attitude to continuously learn and explore new areas of research has influenced me and would greatly benefit me in my future career.

I am grateful to DSO National Laboratories for awarding me the DSO Postgraduate Scholarship to fund my PhD research. My sincere thanks also go to Dr. Alvina Goh, Dr. Pang Sze Kim and Prof Lee Chin-Hui for their advice, encouragement, and efforts that made it possible for me to obtain the scholarship.

I would also like to thank the members of the Distributed and Adaptive Algorithms in Sensor Networks (DASN) group. My sincere appreciation goes to Dr. Ji Feng, Dr. Pratibha Rana, Tang Wenchang, Yang Jielong, Kang Qiyu whom I have worked together with for the insightful discussions and their friendship.

To my friends, Wenjie, Cheng Ta, Ericsson, Yan Jun ,and Shao Xuan, I am grateful for the occasional meet-ups for me to recharge and power through this journey.

Last but not least, I would like to thank my family: my parents, my wife ,and daughters for supporting me throughout this journey.

To my dear family

Contents

Statement of Originality	i
Supervisor Declaration Statement	ii
Authorship Attribution Statement	iii
Acknowledgements	v
Table of Contents	vii
Abstract	x
List of Figures	xi
List of Tables	xiv
List of Abbreviations	xv
List of Symbols	xvii
1 Introduction	1
1.1 Background	1
1.2 Research Motivation and Related Work	3
1.2.1 QCD with unknown post-change distribution	3
1.2.2 QCD in the presence of a nuisance change	5
1.2.3 QCD with sampling constraints	7
1.2.4 QCD with privacy considerations	9
1.3 Thesis Contributions and Organization	11
2 Preliminaries	15
2.1 Probability	15
2.2 Sequential Hypothesis Testing	18
2.3 Quickest Change Detection	19
3 QCD with unknown post-change distribution	22
3.1 Problem formulation	23

3.2	Test statistic based on binning	25
3.3	Properties of the BG-CuSum statistic	29
3.3.1	Estimating a lower bound for the ARL	29
3.3.2	Growth rate of the BG-CuSum statistic	29
3.4	Numerical Results	31
3.4.1	Synthetic data	31
3.4.2	Choice of N	36
3.4.3	Asymptotic behaviour of the ARL and ADD of $\tilde{\tau}$	36
3.4.4	WISDM Actitracker Dataset	38
Appendix 3.A	Proof of Theorem 3.1	40
Appendix 3.B	Proofs of Results in Section 3.3	43
3.B.1	Proof of Proposition 3.1	43
3.B.2	Proof of Proposition 3.2	45
3.B.3	Proof of Proposition 3.3	47
3.B.4	Proof of Proposition 3.4	49
3.B.5	Proof of Proposition 3.5	51
Appendix 3.C	Characterization of the number of bins N	52
3.C.1	Properties of $D(\mathbb{P}^f, N)$	52
3.C.2	Proof of Theorem 3.C.1	56
3.C.3	Simulations	61
4	QCD in the presence of a nuisance change	63
4.1	Problem formulation	63
4.2	Test Statistic for QCD with Nuisance Change	67
4.2.1	Log Likelihood Ratio Growth Rates	71
4.2.2	Conditions for Asymptotic Optimality	73
4.3	Parametrized Families of Post-Change Distributions	77
4.4	Numerical Results	80
4.4.1	W-SGLR on Synthetic Data Satisfying Assumption 4.2	80
4.4.2	W-SGLR on Synthetic Data Violating Assumption 4.2	85
4.4.3	Parametrized Post-Change Distributions	87
4.4.4	Real Data	90
Appendix 4.A	Proof of Propositions 4.1 and 4.2	94
4.A.1	Proof of Proposition 4.1	99
4.A.2	Proof of Proposition 4.2	100
Appendix 4.B	Proof of Lemma 4.1	101
Appendix 4.C	Proof of Lemma 4.2	102
Appendix 4.D	Proof of Proposition 4.3	104
Appendix 4.E	Proof of Theorem 4.3	105
Appendix 4.F	Proof of Lemma 4.4	106
5	QCD with sampling constraints	109
5.1	Problem formulation	109

5.2	Properties of the AOSC	112
5.3	Structure of Asymptotically Optimal Stopping time	115
5.3.1	Asymptotic Properties of GLR CuSum	115
5.4	Optimal Sampling Policy	118
5.4.1	Equal Switching Costs	118
5.4.2	Unequal Switching Costs	119
5.4.2.1	Window size $W = 0$	120
5.4.2.2	Window size $W = 1$	123
5.5	Numerical Results	126
Appendix 5.A	Proof of Lemma 5.1	131
Appendix 5.B	Proof of Lemma 5.2	133
Appendix 5.C	Proof of Lemma 5.3	134
Appendix 5.D	Proof of Proposition 5.2	135
Appendix 5.E	Proof of Theorem 5.1	136
Appendix 5.F	Proof of Proposition 5.3	136
Appendix 5.G	Proof of Proposition 5.4	137
6	Privacy-Aware Quickest Change Detection	140
6.1	Problem formulation	140
6.2	Asymptotic optimality	144
6.3	Relaxation of the Channel Design Problem	145
6.3.1	Relaxation of the Objective Function	145
6.3.2	Relaxation of Privacy Constraints	146
6.3.2.1	Maximal Leakage privacy	146
6.3.2.2	Relaxation of the sequential hypothesis testing privacy constraint	149
6.4	Algorithms	150
6.4.1	Maximal Leakage Privacy	151
6.4.1.1	Exact Method	151
6.4.1.2	Augmented Lagrangian Method	152
6.4.2	Sequential Hypothesis Testing Privacy	153
6.5	Numerical Experiments	154
6.5.1	Maximal Leakage Privacy	154
6.5.2	Sequential Hypothesis Testing Privacy	156
7	Conclusion and Future Work	158
	List of Author's Publications	162
	Bibliography	164

Abstract

Quickest change detection (QCD) is the problem of sequentially detecting a change in the statistical properties of a signal. As sensor and computing technology becomes increasingly pervasive and ubiquitous, QCD has found a wide spectrum of applications, from fraud detection to power system line outage detection. In many of these applications, rather than having one distribution, it is more likely that one of multiple possible distributions generate the signal in the post-change regime. Furthermore, there are additional constraints due to operational requirements that need to be satisfied in addition to traditional QCD constraints. The research goal of this thesis is to study these problems and develop algorithms for QCD under some of these operational constraints.

We first consider the problem of QCD when modeling or estimating the post-change distribution is impractical. A sequential change detection procedure that partitions the sample space into a finite number of bins and detects the change by observing the number of samples that fall into each of these bins is proposed. We developed an algorithm to update the test statistic recursively. Asymptotic properties of the test statistics are derived to offer intuition into its performance trade-off between average run length to a false alarm and the worse case average detection delay. Numerical experiments on synthetic and real data suggest that our proposed method is comparable or better in performance to existing non-parametric change detection methods.

Next, we study the situation where the signal may undergo two types of change, nuisance and critical. We consider the problem where we are only interested in detecting the critical change as quickly as possible while not raising a false alarm when there is no change or when a nuisance has taken place. We develop a sequential change detection procedure based on the generalized likelihood ratio (GLR) test statistic and derive a recursive update scheme for this procedure. Our proposed stopping time is shown to be asymptotically optimal in the sense of Lorden, where it is not possible to improve the trade-off between the logarithm of the average

run length and the worst case average detection delay, under mild technical conditions. Furthermore, when the post-change distribution belongs to a parameterized family, we proposed a generalized stopping time and provided a lower bound on its average run length. We show that our proposed stopping time outperforms the finite moving average stopping time and the naive 2-stage stopping procedure via experiments on real and synthetic datasets.

Another operational constraint that we study is sampling constraints. We consider the problem of QCD for a signal that is not fully observable and can only be observed using a finite set of actions. We propose a static sampling policy and a stopping rule based on the sampling policy. We show that the GLR CuSum achieves asymptotic optimality with a properly designed sampling policy and formulate an optimization problem for which the sampling policy that achieves asymptotic optimality is a solution. Finally, simulation results on partially observed graph signals show that our proposed approach achieves a much lower average detection delay, compared to a random uniform sampling policy, for large average run lengths.

Finally, we consider the problem of QCD while taking privacy considerations into account. We formulate the privacy-aware QCD problem by including a privacy constraint to Lorden's minimax formulation. We show that the GLR CuSum achieves asymptotic optimality with a properly designed sanitization channel and formulate the design of this sanitization channel as an optimization problem. We develop relaxations for the channel design problem to improve its computational tractability.

List of Figures

3.1	Examples of test statistics $\widehat{S}(t)$, $\widetilde{S}(t)$ and $\widehat{S}(t) - \widetilde{S}(t)$ as a function of t for $D(g_N \parallel f_N) = 2$	33
3.2	Examples of test statistics $\widehat{S}(t)$, $\widetilde{S}(t)$ and $\widehat{S}(t) - \widetilde{S}(t)$ as a function of t when the change is in the discrete component such that the pre-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.25\delta_{-1} + 0.25\delta_1$ and post-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.33\delta_{-1} + 0.17\delta_1$	34
3.3	Graph of ADD against $\log(\text{ARL})$ for varying values of N with $f = \mathcal{N}(0, 1)$ and $g = 0.6\mathcal{N}(1, 1) + 0.4\mathcal{N}(-1, 1)$	35
3.4	Graph of ADD against $\log(\text{ARL})$ for varying values of N with $f = \mathcal{N}(0, 1)$ and $g = \mathcal{N}(1, 1)$	35
3.5	Plot of $\Delta(b)$ against threshold b	37
3.6	Comparison of the trade-off between ADD and $\log(\text{ARL})$ for $\widetilde{\tau}$ and $\widehat{\tau}$	38
3.7	Examples of trial performed for pre-change activity of walking, and post-change activity of (a) jogging and (b) ascending upstairs. The black dotted line indicates the boundary between the pre-change and post-change regimes.	39
3.8	Box-plot of the detection delays of each algorithm when the ARL is 6000.	40
3.9	(a) Graph of the upper and lower bounds computed in Algorithm 1. (b) Comparison of ADD vs ARL performance using $N = N_{\text{exact}}$ and $N = N_{\text{approx}}$	62
4.1	Evolution of the signal distribution in our model. In this example, changes in mean are considered nuisance and changes in variance are considered critical.	65
4.2	W-SGLR test statistics $S_{\text{W-SGLR}}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(2, 10)$ and $I = 3.34$	81
4.3	Comparison of the trade-off performance for $\tau_{\text{W-SGLR}}$, τ_{GLRT} , τ_{FMA} and $\tau_{2\text{-stage}}$	82
4.4	W-SGLR test statistics $S_{\text{W-SGLR}}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(5, 20)$ and $I = 3.34$	84
4.5	Comparison of the trade-off performance for $\tau_{\text{W-SGLR}}$, $\tau_{2\text{-stage}}$ and τ_{FMA} when Assumption 4.2 is violated with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 5)$, $g = \mathcal{N}(3, 10)$ and $g_n = \mathcal{N}(5, 10)$	86

4.6	The generalized W-SGLR test statistics $\hat{S}_{\text{W-SGLR}}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(2, 10)$ and $I = 3.34$	88
4.7	Comparison of trade-off performance for the proposed stopping time $\hat{\tau}_{\text{W-SGLR}}$ and $\tau_{\text{W-SGLR}}$	90
4.8	Plots of the pdfs of the observed de-trended signal in different scenarios.	91
4.9	Examples of the W-SGLR test statistic with $m_b = 1024$	92
4.10	Trade-off between the empirical ADD and ARL.	95
5.1	Graph $\mathcal{G}$ generated using the Erdős-Rényi (ER) random graph model with $N = 6$ nodes and probability of an edge $p = 0.2$	128
5.2	Comparison of the ARL-WADD trade-off rate for different stopping times when $\mathcal{C} = \mathcal{C}^2$	130
5.3	Comparison of the ARL-WADD trade-off rate for different stopping times when $\mathcal{C} = \mathcal{C}^3$	130
6.1	Comparison of trade-off between the privacy budget ϵ and the expected KL divergence $\mathbb{E}[D(T_q g_n \ T_q f)]$ for the Exact and augmented Lagrangian method.	155
6.2	Comparison of the time taken to solve the relaxed channel design problem for the Exact and augmented Lagrangian method.	156
6.3	Comparison of trade-off between the privacy ϵ_1 of G_1 in G and the expected KL divergence for different distinguishability ϵ_2 of G_2 in G	157
6.4	Comparison of trade-off between the distinguishability ϵ_2 of G_2 in G and the expected KL divergence for different privacy ϵ_1 of G_1 in G	157

List of Tables

3.1	ADD for post-change distribution with pdf $\mathcal{N}(\delta, 1)$ with ARL = 500.	32
3.2	ADD for post-change distribution with pdf $\mathcal{N}(0, \delta^2)$ with ARL = 500.	33
3.3	ADD for Laplace post-change distribution with ARL = 500.	33

List of Abbreviations

ADD	Average Detection Delay
AOSC	Average Observation-Switching Cost
ARL	Average Run Length
BG-CuSum	Binned Generalized CuSum
CuSum	Cumulative Sum
EWADD	Expected Worst-Case Average Detection Delay
FMA	Finite Moving Average
GLR	Generalized Likelihood Ratio
GLRT	Generalized Likelihood Ratio Test
HMM	Hidden Markov Model
i.i.d.	independent and identically distributed
IoT	Internet of Things
IRM	iterative rank minimization
KL	Kullback-Leibler
LHS	left-hand side
MMSE	Minimum Mean-Square Error
pdf	probability density function
QCD	Quickest Change Detection
QCQP	Quadratically Constrained Quadratic Program
RHS	right-hand side
s.t.	subject to
SDR	Sequential Decision Rule
SGLR	Simplified Generalized Likelihood Ratio

SPRT	Sequential probability ratio test
SRP	Shiryayev-Roberts-Pollak
TCD	Transient Change Detection
w.r.t.	with respect to
WADD	Worst-case Average Detection Delay
W-SGLR	Window-limited Simplified Generalized Likelihood Ratio

List of Symbols

Throughout this thesis, we use upper-case (e.g. X) to refer to a random variable, and lower case (e.g. x) to refer to a realization of the random variable X .

$\mathbb{R}$	The set of real numbers
$\mathbb{R}_{>0}$	The set of positive real numbers
$\mathbb{N}$	The set of positive integers
$\mathbb{E}^f$	The mathematical expectation w.r.t. the probability distribution with generalized pdf f
$\mathbb{P}^f$	Probability measure with generalized pdf f
$X \sim f$	The random variable X has distribution with generalized pdf f
$\mathcal{N}(\mu, \sigma^2)$	The pdf of the Gaussian distribution with mean μ and variance σ^2
δ_θ	The Dirac delta function at θ
$ A $	The cardinality of the set A
$\mathbf{1}_A$	The indicator function of the set A
$D(g \parallel f)$	The KL divergence of distribution with pdf f from distribution with pdf g
$\xrightarrow{\mathbb{P}}$	The convergence in $\mathbb{P}$ -probability
$\xrightarrow[\text{a.s.}]{\mathbb{P}}$	Almost-sure convergence under the probability measure $\mathbb{P}$
$\text{supp}(f)$	The support of a probability mass function f
$\alpha[j]$	The j -th entry of a vector $\alpha = (a_1, a_2, \dots, a_n)$
$T[i, j]$	The probability of moving to a state j given that it is currently at state i for a probability transition matrix T
ess sup	The essential supremum
$(x)^+$	$\max\{x, 0\}$

Chapter 1

Introduction

1.1 Background

Quickest change detection (QCD) is the statistical problem of detection for a variation in the statistical properties of a signal with minimal delay. Under a simple QCD scenario, given a sequence of independent and identically distributed (i.i.d.) observations $\{x_t : t \in \mathbb{N}\}$ generated with a distribution $\mathbb{P}^f$ up to $t = \nu - 1$ and generated i.i.d. with a distribution $\mathbb{P}^g \neq \mathbb{P}^f$ after the change point $t = \nu$, the aim is to detect this change as quickly as possible while maintaining a false alarm constraint.

Under the Bayesian interpretation of QCD, the change-point is assumed to be a random variable with a prior distribution. When this prior distribution is a geometric distribution with both the pre- and post-change distribution being known, Shiryaev [1–3] proposed a detection procedure based on comparing the posterior probability of the change currently in effect against a fixed detection threshold. Shiryaev also proved the optimality of this detection procedure. The authors of [4, 5] developed an asymptotic Bayesian theory of QCD for a class of non-i.i.d. signal models and prior distributions.

For the non-Bayesian formulation of QCD, we assume that the change-point is unknown but deterministic. Page [6] developed the Cumulative Sum (CuSum) control chart for QCD when both the pre- and post-change distributions are known. The CuSum test is shown to be asymptotically optimal, as the false alarm rate goes to zero, by Lorden [7]. It was later shown by Moustakides [8] that for i.i.d. signal models, under Lorden's optimality criterion, the CuSum test is exactly optimal. Lai [9] later proved that the CuSum test is asymptotically optimal under Pollak's criterion [10], as the false alarm rate goes to zero and further extends to the case when the data are dependent. For the case of finite multiple post-change distributions, Lorden [7] showed that the generalized likelihood-ratio (GLR) CuSum is asymptotically optimal. The authors [9, 11–15] also considered the QCD problem when the post-change distribution is unknown to a certain degree. We refer the reader to [16–18] and the references therein for a comprehensive overview of the QCD problem.

Traditionally, the problem of QCD has applications in quality control for manufacturing [19, 20] where it is essential that we quickly detect any deviation in product quality. With the proliferation of sensor and computing technologies, QCD has found new applications in areas such as structural health monitoring [21], spam detection [22–24], fraud detection [25], cognitive radio [26], bioinformatics [27], power system line outage detection [28], sensor networks [29–31], and network surveillance [32–34]. As each of these applications has different requirements, it is ever more important to study QCD under each of the specific operational constraints. In many of these applications, the typical assumptions of traditional QCD problems may not be satisfied. Furthermore, additional constraints may need to be included when formulating the QCD problem to take application-specific considerations into account. In the application of bearing failure detection [35], rather than having one type of change, the signal of interest undergoes several types of changes with only a subset of these changes that need to be quickly detected while the rest of the changes are to be ignored. In the application of physical activity change detection [36], rather than having a known post-change distribution, the

post-change activity generating the signal is completely unknown and it will be impractical to attempt to model such a post-change distribution. In applications where the signal is obtained from a sensor network with a bandwidth constraint, the bandwidth constraint introduces new challenges to the tradition QCD formulation. In this thesis, we consider the non-Bayesian QCD problem under some of these operational constraints.

1.2 Research Motivation and Related Work

In this section, we highlight the research motivations of the QCD problem under the different operational constraints and their related works.

1.2.1 QCD with unknown post-change distribution

The first operational constraint we consider in this thesis is the QCD problem with a known pre-change distribution and an unknown post-change distribution. The assumption that the pre-change distribution is known is reasonable because, in most practical applications, a large amount of data generated by the pre-change distribution is available to the observer who may use this data to obtain an accurate approximation of the pre-change distribution [37, 38]. However, estimating or even modeling the post-change distribution is often impractical as we may not know *a priori* what kind of change will happen. Furthermore, in many applications such as structural failure detection, data generated by the post-change distribution is either difficult or expensive to obtain. Hence, we assume that the only information we have about the post-change distribution is that it belongs to a set of distributions that is distinguishable from the pre-change distribution when the sample space is discretized into N bins. We seek to design a low-complexity detection algorithm that allows us to quickly detect the change, under false alarm constraints, and with minimal knowledge of the post-change distribution.

There are many existing works that consider the QCD problem where the post-change distribution is unknown to a certain degree. In [11], the authors considered the case where the post-change distribution belongs to a one-parameter exponential family with the pre-change distribution being known. The case where both the pre- and post-change distributions belong to a one-parameter exponential family was considered by [9]. In [12], the authors developed a data-efficient scheme that allows for optional sampling of the observations in the case when either the post-change family of distributions is finite, or both the pre- and post-change distributions belong to a one-parameter exponential family. In [39], the authors assumed that the observations are time-dependent as in ARMA, general linear processes and α -mixing processes. In [40], the authors proposed using an infinite hidden Markov model for tracking parametric changes in the signal model. Classical approaches to the QCD problem without strong distributional assumptions can be found in [41–44]. Although there are no distributional assumptions, the type of change assumed in [41–45] is a shift in the mean and in [46], a shift in the scale of the observations. In [47], the authors provided a kernel-based detection scheme for a change-point detection problem where the post-change distribution is completely unknown. However, the kernel-based detection scheme requires choosing a proper kernel bandwidth. This is usually done by tuning the bandwidth parameter on a reference dataset representative of the pre-change and possible post-change conditions. In [48], the authors proposed a non-parametric algorithm for detecting a change in mean. They assume that the mean of the pre- and post-change distributions are not equal but do not assume any knowledge of the distribution function.

For our work under this operational constraint, we consider the case where less information is available about the post-change distribution as compared to [9, 11, 12]. The only assumption we make is that the post-change distribution belongs to a set of distributions that are distinguishable from the pre-change distribution when the sample space is discretized into a known number of bins. Our method is also more general than [41–46, 48] as it is not restricted to a shift in mean, location or scale.

Related work is discussed in [49], in which an asymptotically optimal universal scheme is developed to isolate an outlier data stream that experiences a change, from a pool of at least two typical streams. Here, we only have a single stream of observations and therefore our scheme is unable to make use of comparisons across streams to detect for changes. Other related works include [42, 47, 50], which do not require prior information about the pre-change distribution. These works first partition the observations using a hypothesized change point into two sample sets and then test if they are generated by the same distribution using a two-sample test. This approach results in weak detection performances when the change occurs early in the observation sequence. For our work under this operational constraint, we assume that prior information about the pre-change distribution is available, which allows us to detect the change even when the change occurs early in the observations.

1.2.2 QCD in the presence of a nuisance change

Another operational constraint we study in this thesis is the QCD problem when the signal of interest may undergo different types of change. However, only a subset of these changes may be of interest to the user. One example is the problem of bearing failure detection using accelerometer readings [35]. During normal operations, the bearings are driven at two different activity levels, idle or active. In a typical bearing failure detection scenario, the bearing is initially driven at the idle state. A change in the activity level results in a change in the statistical properties of the accelerometer readings. However, this change is not of interest to us and is called a nuisance change. We are only interested in the change arising from the failure of the bearing, which is known as a critical change. Furthermore, the statistical properties of the observations obtained when the bearing is faulty depend on the activity level that it is driven at. The traditional QCD framework does not allow us to distinguish between critical and nuisance changes. Furthermore, due to the nuisance change, the observations are no longer i.i.d. either in the pre-change or post-change regime, depending on when the nuisance change occurs. Under this

operational constraint, we investigate the non-Bayesian formulation of the QCD problem under a nuisance change and propose a window-limited stopping time that ignores the nuisance change but detects the critical change as quickly as possible.

Existing works in QCD that consider the problem where observations are not generated i.i.d. before and after the change-point can be categorized into three main categories. In the first category, the papers [51–53] consider the problem where the pre-change distribution and the post-change distribution are modeled as hidden Markov models (HMMs). In [51], the authors proved the asymptotic optimality of the CuSum procedure for the HMM signal model in the sense of Lorden. In [52], the authors developed the Shiriyayev-Roberts-Pollak (SRP) rule for the HMM signal model and proved its optimality in the sense of Pollak. The authors of [53] consider the problem where the vector parameter of a two-state HMM changes at some unknown time. The second category of papers [5, 54] considers a QCD problem which relaxes the i.i.d. assumption. In [54], the authors established the optimality of CuSum and the Shiriyayev-Roberts stopping rule in the class of random processes with likelihood ratios that satisfy certain independence and stationary conditions. The class of random processes includes Markov chains, AR processes, and processes evolving on a circle. In [5], the authors considered the Bayesian QCD problem where conditions on the asymptotic behavior of the likelihood process are assumed. Unlike all the aforementioned papers, the signal model in our QCD problem with nuisance change cannot be modeled by an HMM, and the likelihood ratios generated by our signal model are non-stationary. In the third category, the papers [55–60] consider QCD of transient changes, where the change is either not persistent or multiple changes occur throughout the monitoring process. In our problem considered under this operational constraint, we allow some changes to be considered nuisance whereas all changes are considered critical in the aforementioned papers.

1.2.3 QCD with sampling constraints

The third operational constraint we study in this thesis is the QCD problem with sampling constraints. In many applications, the signal of interest x_t may be high dimensional. For example, x_t may consist of observations from many correlated sensors. Due to the large number of sensors in the network, bandwidth and power constraints prevent us from observing the entire network at any time instance, and we may only obtain sensor readings from a small subset of sensors at any time instance [61, 62]. While it may seem optimal to observe the maximum number of sensors allowed by the network, this sampling policy may not be feasible due to power and communication bandwidth considerations. Furthermore, the action of switching from one subset to another subset of sensors also incurs power and communication costs. Under this operational constraint, we consider both of these costs collectively as the *observation-switching cost*, and we study the problem of QCD while maintaining an average observation-switching cost (AOSC) constraint. To be more general, we consider the case where the signal can only be observed using an action selected from a set of permissible actions with observation-switching costs associated with the sequence of actions chosen. We assume that the pre- and post-change distributions, as well as their conditional distributions given the actions, are known to the observer. The objective is to design a sampling policy together with a stopping time that satisfies both the QCD false alarm and AOSC requirements. To solve this problem, we propose a sampling policy that determines which action to perform based on the previous M actions and a generalized likelihood ratio (GLR) Cumulative Sum (CuSum) stopping time for the QCD problem.

Existing works in QCD where the signal is not entirely available to the decision-maker or the fusion center can be categorized into three main categories. In the first category, the papers [63–66] consider the problem of distributed or decentralized QCD where each node observes and processes its signal locally, with some memory of its previous messages, before sending a message to the fusion center. The authors of [64] consider the problem where each sensor only has access to the

local information at that node and would process the signal to send a quantized message to the fusion center for further processing.

The second category of papers [12, 67–69] considers a QCD problem where there is an additional cost for sensing at each node and a control policy that determines whether a given observation is taken. In [12], the authors developed a data-efficient scheme that allows for optional on-off sampling of the observations in the case where either the post-change family of distributions is finite, or both the pre- and post-change distribution belong to a one-parameter exponential family.

In the third category, the papers [70–73] consider QCD where the observer only has access to compressed or incomplete measurements. The authors of [70] study the problem of sequential change-point detection where a randomly generated linear projection is used to reduce the dimensions of a high dimensional signal for the purpose of QCD. In [73], the authors propose to monitor the top- r nodes with most significant local CuSum and introduce a compensating coefficient to the CuSum of the nodes not making an observation for the purpose of QCD when only a subset of the nodes are observable at any time instance. In [72], the authors consider QCD with closed-loop control of the actions where the current nodes to observe is determined by the maximal likelihood estimate of the post-change hypothesis. In [14], we discussed the QCD problem where the observer is only able to obtain a partial observation of the signal through an action with an open-loop control of the actions.

Unlike the papers mentioned above, we provide a general framework by considering random decision rules to select the current actions. We also do not give a fixed cost to the sampling of observation. Instead, we consider a more general approach where we use a set of permissible actions to model the practical sampling constraints and a cost is associated with the switching of actions to model observation-switching costs. Under this operational constraint, we consider the case where the decision-maker is given a finite set of pre-defined actions, and the observed signal is a

function of the action and the full signal. We also do not make any assumptions about the pre- and post-change distributions.

1.2.4 QCD with privacy considerations

The fourth operational constraint we study in this thesis is the QCD problem with privacy considerations. There are two potential types of applications of QCD which require privacy considerations. In the first type of application, rather than having one distribution, the distribution which generates the signal in the post-change regime belongs to a set $G = \{\mathbb{P}^{g_1}, \dots, \mathbb{P}^{g_{|G|}}\}$. The distribution of some sensitive information U may depend on the post-change distribution. Thus, knowing the distribution $\mathbb{P}^g \in G$ which generates the signal in the post-change regime may reveal some information about U . One example of such an application is occupancy detection. In an Internet of Things (IoT) based occupancy detection system, the occupancy sensor continuously measures attributes such as infrared radiation, temperature, humidity and carbon dioxide levels, to quickly detect when a room becomes occupied so that certain functions like turning on the ceiling lights can be automated. A baseline distribution can be used to model the fluctuation of these attributes with time when the room is empty. However, fluctuations in these attributes when the room is occupied can reveal the number of people and the activity conducted in the room, leading to privacy leakage if an adversary has access to the raw attribute signals. Our goal is to sanitize the attribute signals so that the change from a vacant to an occupied room can still be detected while preserving the privacy of the occupants in the room. In practice, we are unable to know the adversary's intent and hence would like to sanitize the signal so that the largest improvement an adversary can achieve, over all possible queries, is controlled.

In the second type of application, G may be partitioned into two sets, a private set G_1 and a public set G_2 . We would like to sanitize the signal so that it is difficult for an adversary to deduce the distribution when a post-change distribution from the private set G_1 is generating the signal. On the other hand, we would like to

retain the ability to deduce the distribution when a post-change distribution from the public set G_2 is generating the signal. One example of such an application is activity monitoring using wearables. In this application, we would like to quickly detect any change from a resting state to an active state. There are many possible active states such as walking, running and typing on a computer. We would like to protect the privacy of some of these active states, like typing on a computer or using a mobile phone, while still being able to accurately track any change in the activity. For the rest of the active states, we would also like to maintain the distinguishability of the signals for potential applications in the future. Hence, we would like to perform QCD while protecting the privacy of some active states and maintaining some distinguishability for the rest of the active states.

As sensor and computing technology become increasingly ubiquitous and powerful, it becomes easier for an adversary to infer sensitive information, such as lifestyle preferences and location information, from the available data and this information becomes increasingly at risk of abuse. Existing work on protecting or quantifying privacy can be divided into two main categories: *data* privacy and *inference* privacy [74–77]. Data privacy refers to the protection of the sensors’ raw information from being obtained by the fusion center. In contrast, inference privacy refers to the protection from an adversary’s attempt to deduce properties of the underlying distribution. Privacy metrics proposed to quantify data privacy include local differential privacy [78–81], k -anonymity [82] and homomorphic encryption [83–85]. On the other hand, privacy metrics proposed to quantify inference privacy include average information leakage [86–88], mutual information privacy [89, 90], maximal leakage privacy [91, 92] differential privacy [93], hypothesis testing adversary privacy [94], compressive privacy [95, 96] and information privacy [97–101]. We refer the reader to [76, 89] for a comprehensive discussion on the relationship between data and inference privacy. Under this operational constraint, we use maximal leakage privacy, and sequential hypothesis testing adversary privacy which is the sequential analog to the privacy metric proposed in [94], to quantify the gain which the adversary obtains through the observation of the signal.

Under this operational constraint, we address the QCD problem with multiple post-change distributions while maintaining a privacy constraint for two different privacy metrics. For each of the privacy metrics, we propose a signal sanitization algorithm and stopping time that is able to identify the critical change quickly while preserving a pre-determined level of privacy.

There are several works in the literature that present privacy-preserving frameworks for different signal processing tasks [101–104]. In [104], the authors developed differentially private algorithms, which assume that the adversary knows all entries of the database except one, for the purpose of change-point detection. Unlike [104], we consider a weaker form of privacy as this assumption may be too strong for some applications. Furthermore, we provide theoretical guarantees on the average run length to a false alarm and the worst-case detection delay which are more relevant to the QCD task as compared to the guarantees on the accuracy of the estimated change-point provided by [104] which are more suited for the change-point detection task. In [102], the sanitization channel is designed for general signal processing tasks while preserving privacy, where the utility of the sanitized signal is measured using general information-theoretic quantities such as mutual information, minimum mean-square error (MMSE), and the probability of error. In [101, 103], the authors consider a fixed sample size problem of hypothesis testing while preserving privacy, where the utility of the sanitized signal is measured by the Bayes error or the type-II error of the test. Here, we consider the problem of optimizing QCD performance while preserving privacy. Furthermore, unlike the papers mentioned above, we consider sequential signals. It is thus possible for the adversary to obtain an arbitrarily large number of samples to improve his guess.

1.3 Thesis Contributions and Organization

In Chapter 3, we consider the QCD problem where the pre-change distribution is known, and the post-change distribution belongs to a family of distributions distinguishable from the pre-change distribution when the sample space is discretized

into bins. Our goal is to develop an algorithm that is effective and has low computational complexity. Our main contributions are as follows:

1. We propose a Binned Generalized CuSum (BG-CuSum) test and derive a recursive update formula for our test statistic, which allows for an efficient implementation of the test.
2. We present properties of the family of post-change distributions distinguishable from the pre-change distribution and indicate how the number of bins can be chosen.
3. We derive a lower bound for the average run length (ARL) to false alarm of our BG-CuSum test, and show asymptotic properties of the BG-CuSum statistic to provide insights into the trade-off between the average detection delay (ADD) and ARL.
4. We provide simulations and experimental results, which indicate that our proposed BG-CuSum test outperforms various other non-parametric change detection methods and its performance approaches a test with known asymptotic properties as the ARL becomes large.

In Chapter 4, we consider the non-Bayesian QCD problem where both nuisance and critical changes may occur, and our objective is to detect the critical change as quickly as possible while ignoring the nuisance change. Our goal is to develop a sequential algorithm with computational complexity that increases linearly with the number of samples observed. Our main contributions are as follows:

1. We formulate the QCD problem with a nuisance change and propose a window-limited simplified GLR (W-SGLR) stopping time.
2. We derive a lower bound for the average run length (ARL) to a false alarm, and the asymptotic upper bound of the worst-case average detection delay (WADD) for our proposed test.

3. We prove the asymptotic optimality of the W-SGLR stopping time under mild technical assumptions.
4. We provide simulation and experimental results that verify the theoretical guarantees of our proposed test and also illustrate the performance of our proposed test on a real dataset.

In Chapter 5, we consider the problem QCD while maintaining an AOSC constraint. The objective is to design a sampling policy together with a stopping time that satisfies both the QCD false alarm ARL and AOSC requirements. Our main contributions are as follows:

1. We formulate the QCD problem with an AOSC constraint with sampling policy of window size $W < \infty$, where the current action is allowed to be dependent only on the previous W actions. We show that the GLR CuSum stopping time together with a properly designed sampling policy is asymptotically optimal. We formulate the design of the sampling policy of window size W as a quadratic programming problem with an additional combinatorial constraint.
2. We derive expressions for the AOSC and the asymptotic WADD for the GLR CuSum stopping time together with a sampling policy of window size $W < \infty$. Using these expressions, we prove that under the constraint that $W \geq 1$, there exists an asymptotically optimal sampling policy of window size $W = 1$.
3. For the case when the window size of the sampling policy is $W = 0$, we provide derivations to transform the policy design problem into a linear programming problem for the special case where all observation-switching costs are equal and also an iterative rank minimization (IRM) algorithm to obtain a locally optimal solution for the general case.

4. For the case when the window size of the sampling policy is $W = 1$, we provide relaxations and derivations to transform the policy design problem into a convex programming problem.

In Chapter 6, we consider the QCD problem with multiple post-change distribution while maintaining a privacy constraint. The objective is to propose a signal sanitization algorithm and stopping time that can identify the critical change quickly while preserving a pre-determined level of privacy. Our main contributions are as follows

1. We formulate the QCD problem with privacy constraints and show that the GLR CuSum stopping time together with a properly designed sanitization channel is asymptotically optimal.
2. For the maximal leakage privacy metric, we propose relaxations to the channel design problem and present methods to solve the relaxed channel design problem exactly and approximately using the augmented Lagrangian method.
3. For the sequential hypothesis testing privacy metric, we propose relaxations to the channel design problem and show that the relaxed channel design problem can be expressed as a mixed-integer linear program.

Chapter 2

Preliminaries

In this chapter, we introduce the preliminary concepts that are necessary for the rest of this thesis. We will briefly discuss probability theory, sequential hypothesis testing and QCD. For a comprehensive treatment of these topics, we refer the interested reader to Durrett [105], Tartakovsky [16] and Poor [18].

2.1 Probability

The concept of sequential information and decision rules is central in QCD. In this section, we will define several concepts used in probability and use them to formalize the concept of sequential information and decision rules. First, we begin by defining the notion of a σ -algebra, **probability space** and **probability measure**.

Definition 2.1. A σ -algebra $\mathcal{F}$ of Ω is a non-empty collection of subsets of Ω such that

- $\emptyset, \Omega \in \mathcal{F}$,
- if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$,
- if $\{A_i\}_{i \in \mathbb{Z}} \subseteq \mathcal{F}$ is a countable sequence of sets then $\cup_{i \in \mathbb{Z}} A_i \in \mathcal{F}$.

A **probability measure** on a σ -algebra $\mathcal{F}$ is a non-negative countably additive set function $\mu : \mathcal{F} \rightarrow \mathbb{R}$ such that

- $\mu(\emptyset) = 0$ and $\mu(A) \geq 0$ for any $A \in \mathcal{F}$,
- if $\{A_i\}_{i \in \mathbb{Z}} \subseteq \mathcal{F}$ is a countable sequence of disjoint sets then

$$\mu \left(\bigcup_{i \in \mathbb{Z}} A_i \right) = \sum_{i \in \mathbb{Z}} \mu(A_i),$$

- we have $\mu(\Omega) = 1$.

A **probability space** is a triple $(\Omega, \mathcal{F}, \mu)$ where Ω is the set of outcomes, $\mathcal{F}$ is a σ -algebra of events such that $A \subseteq \Omega$ for any $A \in \mathcal{F}$ and $\mu : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure.

In applications, we must assign numerical values to events so that we can perform computations such as averages. This can be formalized using the concept of **random variables**.

Definition 2.2. Given a probability space $(\Omega, \mathcal{F}, \mu)$, a $\mathbb{R}^N$ -valued function X defined on Ω is said to be a **random variable** if and only if for every Borel set $B \subseteq \mathbb{R}^N$, we have

$$X^{-1}(B) \triangleq \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

A **random variable** X is said to be **measurable with respect to** a σ -algebra $\mathcal{G} \subseteq \mathcal{F}$ if and only if for every Borel set $B \subseteq \mathbb{R}^N$, we have

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{G}.$$

Furthermore, each random variable X induces a **probability space** $(\mathbb{R}^N, \mathcal{B}, f)$ with

$$f(B) = \mu(X^{-1}(B)) \quad \text{for any borel set } B \subseteq \mathcal{B}.$$

We call f the **distribution** of the random variable X .

Using techniques [106] from measure theory, we can define the **expectation** of X using the integral

$$\mathbb{E}[X] = \int_{\Omega} X \, df.$$

Finally, in order for us to discuss the concept of obtaining observations sequentially, we define the notions of a filtration and stopping time.

Definition 2.3. A **filtration** of $\mathcal{F}$ is a sequence of sets $\mathcal{F}_k$ such that

- $\mathcal{F}_k \subseteq \mathcal{F}$ is a σ -algebra for all $k \in \mathbb{N}$,
- $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$.

In sequential testing applications, the decision rule would only have access to the observations obtained up to the present. The notion of **stopping time** precisely describes this:

Definition 2.4. A **stopping time** with respect to a filtration $\mathcal{F}$ is an extended random variable

$$\tau : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}$$

such that $\{\omega \in \Omega \mid \tau(\omega) \leq k\} \in \mathcal{F}_k$.

Finally, we present a theorem by Wald [107] that we use for the analysis of stopping times.

Theorem 2.1. If τ is a stopping time with respect to an i.i.d. sequence $\{X_k \mid k = 1, 2, \dots\}$ and if $\mathbb{E}[\tau] < \infty$ and $\mathbb{E}[|X_1|] < \infty$, then

$$\mathbb{E} \left[\sum_{k=1}^{\tau} X_k \right] = \mathbb{E}[\tau] \mathbb{E}[X].$$

2.2 Sequential Hypothesis Testing

In this section, we will formulate the classical sequential detection problem and also introduce the sequential probability ratio test. Finally, we present some important results that will be used in this thesis.

Consider a sequence $\{X_k \mid k = 1, 2, \dots\}$ of independent and identically distributed (i.i.d.) observations such that they are generated by one of the hypotheses:

$$H_0 : X_k \sim f_0, \quad k = 1, 2, \dots$$

$$H_1 : X_k \sim f_1, \quad k = 1, 2, \dots$$

where f_0, f_1 are two distinct distributions on $(\mathbb{R}^N, \mathcal{B})$. We obtain the observations sequentially and would like to decide which hypothesis is generating the observations as soon as possible, subjected to some error constraints. Let $\mathcal{T}$ denote all stopping times τ with respect to the filtration $\{\sigma(X_1, X_2, \dots, X_k) : k \geq 0\}$ where $\sigma(X_1, X_2, \dots, X_k)$ is the smallest σ -algebra with respect to which all $X_1, X_2, \dots, X_k$ are measurable, and $\mathcal{D}$ refer to the set of all terminal decision rules $\delta = \{\delta_k \mid k \in \mathbb{N}\}$ where δ_k is $\sigma(X_1, X_2, \dots, X_k)$ -measurable. The set of all sequential decision rules (SDR) is the pair $(\mathcal{T}, \mathcal{D})$. Any SDR $(\tau, \delta) \in (\mathcal{T}, \mathcal{D})$ can be described as a decision τ for when to stop sampling and δ_τ for deciding which hypothesis to accept.

An example of a sequential detection rule is the **sequential probability ratio test** (SPRT):

For a given $0 < A \leq 1 \leq B < \infty$, the SPRT (τ, δ) with lower and upper threshold

A and B is

$$\begin{aligned}\tau &= \inf\{k \mid \Lambda_k \notin (A, B)\}, \\ \delta_k &= \mathbb{1}_{\Lambda_k \geq B}, \\ \Lambda_k &= \prod_{j=1}^k \frac{df_1}{df_0}(X_j), \Lambda_0 = 1 \quad \text{for } k = 1, 2, 3, \dots,\end{aligned}$$

where $\frac{df_1}{df_0}$ is the Radon-Nikodym derivative of f_1 with respect to f_0 . Finally, we present a theorem by Wald and Wolfowitz [108] which states that the SPRT has the minimum expected stopping time amongst all SDR given the error probabilities:

Theorem 2.2. Suppose (τ, δ) is a SPRT with lower and upper threshold, A and B respectively, where $0 < A \leq 1 \leq B < \infty$ and $\mathbb{P}_1(\delta_\tau = 0) + \mathbb{P}_0(\delta_\tau = 1) \leq 1$. Let (τ', δ') denote any other SDR such that $\max\{\mathbb{E}_0[\tau'], \mathbb{E}_1[\tau']\} < \infty$ and satisfies

$$\mathbb{P}_0(\delta'_\tau = 1) \leq \mathbb{P}_0(\delta_\tau = 1) \quad \text{and} \quad \mathbb{P}_1(\delta'_\tau = 0) \leq \mathbb{P}_1(\delta_\tau = 0),$$

where $\mathbb{E}_i, \mathbb{P}_i$ are the expectation and probability assuming the hypothesis H_i .

Then, we have $\mathbb{E}_0[\tau'] \geq \mathbb{E}_0[\tau]$ and $\mathbb{E}_1[\tau'] \geq \mathbb{E}_1[\tau]$.

2.3 Quickest Change Detection

Given an i.i.d. sequence of observations that is distributed with distribution f_0 up to an unknown change point ν , and is i.i.d. with distribution $f_1 \neq f_0$ after, the goal of QCD is to detect the change in distribution as quickly as possible, subject to false alarm constraints. Formally, QCD can be formulated as follows:

Let f_0 and $f_1 \neq f_0$ be distributions, and $X_1, X_2, \dots$ be a sequence of random variables satisfying the following:

$$\begin{cases} X_t \sim f_0 & \text{i.i.d. for all } t < \nu, \\ X_t \sim f_1 & \text{i.i.d. for all } t \geq \nu, \end{cases} \quad (2.1)$$

where $\nu \geq 0$ is an unknown but deterministic change point. The goal of QCD is to find a stopping time τ^* with respect to the filtration $\{\sigma(X_1, X_2, \dots, X_k) : k \geq 0\}$ that satisfies

$$\tau^* = \arg \min_{\tau \in \mathcal{T}} \text{WADD}(\tau),$$

where $\mathcal{T}$ is the set of stopping times τ with $\text{ARL}(\tau) \geq \gamma$. Here, $\text{WADD}(\tau)$ is the worst case average detection delay and $\text{ARL}(\tau)$ is the average run length of a stopping time τ defined as

$$\begin{aligned} \text{WADD}(\tau) &= \sup_{\nu \geq 1} \text{ess sup } \mathbb{E}_\nu [(\tau - \nu + 1)^+ \mid X_1, \dots, X_{\nu-1}], \\ \text{ARL}(T) &= \mathbb{E}_\infty [\tau], \end{aligned}$$

where $\mathbb{E}_\nu[\cdot]$ denotes the expectation assuming the change point is at ν . This is known as the Lorden's Criterion [7].

For the case when the pre- and post-change distributions are known, Page [6] developed the CuSum test for QCD. Lorden [7] showed the asymptotic optimality of the CuSum test as the false alarm rate goes to zero. We summarize these results in the following theorem:

Theorem 2.3. Let $\{X_k\}_{k \in \mathbb{N}}$ be a random sequence with an unknown but deterministic change point ν . We further assume that $X_1, \dots, X_t$ are i.i.d. with distribution f_0 and $X_t, X_{t+1}, \dots$ are i.i.d with distribution f_1 . Let τ_{CuSum} be the stopping time defined by

$$\begin{aligned} \tau_{\text{CuSum}} &= \inf \left\{ t \mid S_{\text{CuSum}}(t) > \frac{1}{a} \right\}, \\ S_{\text{CuSum}}(t) &= \max_{1 \leq j \leq t+1} \sum_{i=j}^t \log \frac{df_1}{df_0}(X_i). \end{aligned}$$

Then τ_{CuSum} satisfies the following

$$\text{ARL}(\tau_{\text{CuSum}}) = \mathbb{E}_\infty [\tau_{\text{CuSum}}] \geq \frac{1}{a} \quad \text{and}$$

$$\text{WADD}(\tau_{\text{CuSum}}) \leq \frac{|\log a|}{\mathbb{E}_1 \left[\log \frac{df_1}{df_0}(Z_1) \right]} \quad \text{as } a \rightarrow 0.$$

Furthermore, let $n(\gamma) = \inf_{\tau} \text{WADD}(\tau)$, where the infimum is taken over all stopping times τ satisfying $\text{ARL}(\tau) \geq \gamma$. If $\mathbb{E}_1 \left[\log \frac{df_1}{df_0}(X_1) \right] < \infty$, then $n(\gamma) \asymp \frac{\log \gamma}{\mathbb{E}_1 \left[\log \frac{df_1}{df_0}(X_1) \right]}$ as $\gamma \rightarrow \infty$.

Chapter 3

QCD with unknown post-change distribution

In this chapter, we discuss the problem of QCD with an unknown post-change distribution. In Section 3.1, we present the signal model and problem formulation. In Section 3.2, we present our GLRT based QCD procedure and propose an approximation that can be computed efficiently. In Section 3.3, we analyse the asymptotic behavior of our test statistic, and provide a heuristic analysis of the trade-off between the ADD and the ARL to false alarm. In Section 3.4, we present simulation and experimental results to illustrate the performance of our algorithm. In Section 3.C, we present an exact characterization of the number of bins required such that the post-change distribution is distinguishable from pre-change distribution.

Notations: For the rest of this chapter, we let $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ denote the measure and mathematical expectation respectively where the change point is at ν for $\nu \in \mathbb{N}$. In particular, $\mathbb{P}_1$ and $\mathbb{E}_1$ denote the measure and mathematical expectation, respectively, when the change has occurred at time $t = 1$, i.e., all the observations are distributed according to the post-change distribution. Similarly, we let $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ denote the measure and mathematical expectation, respectively, when there is no change, i.e., all the observations are distributed according to the pre-change distribution.

3.1 Problem formulation

We consider distributions defined on the sample space $\mathbb{R}$. Let $f = p_0 f_c + \sum_{h=1}^H p_h \delta_{\theta_h}$ be the generalized pdf of the pre-change distribution $\mathbb{P}^f$, where f_c is the pdf of the continuous part, $\{\theta_h \in \mathbb{R} : h = 1, \dots, H\}$ are the locations of the point masses, δ_{θ} is the generalized pdf of the point mass distribution localized at θ such that

$$X \sim \delta_{\theta} \quad \text{if and only if} \quad \mathbb{P}^{\delta_{\theta}}(X = \theta) = 1$$

and $0 \leq p_h \leq 1$ for all $h \geq 0$ with $\sum_{h=0}^H p_h = 1$. Similarly, let $g = q_0 g_c + \sum_{h=1}^H q_h \delta_{\theta_h}$ be the generalized pdf of the post-change distribution $\mathbb{P}^g$ where $g \neq f$, and $0 \leq q_h \leq 1$ for all $h \geq 0$ with $\sum_{h=0}^H q_h = 1$. Let $X_1, X_2, \dots$ be a sequence of real valued random variables satisfying the following:

$$\begin{cases} X_t \sim f & \text{i.i.d. for all } t < \nu, \\ X_t \sim g & \text{i.i.d. for all } t \geq \nu, \end{cases} \quad (3.1)$$

where $\nu \geq 0$ is an unknown but deterministic change point. The quickest change detection problem is to detect the change in distribution, through observing $X_1 = x_1, X_2 = x_2, \dots$, as quickly as possible while keeping the false alarm rate low. In this chapter, we assume that the observer only knows the pre-change distribution $\mathbb{P}^f$ but does not have full knowledge of the post-change distribution $\mathbb{P}^g$. We assume that the post-change distribution is distinguishable from the pre-change distribution when sample space is discretized into a known number of bins, defined as follows.

Definition 3.1. Let $\Theta = \{\theta_1, \dots, \theta_H\}$. For any $N \geq 1$, let $I_1^N = (-\infty, z_1] \setminus \Theta$, $I_2^N = (z_1, z_2] \setminus \Theta$, $\dots$, and $I_N^N = (z_{N-1}, \infty) \setminus \Theta$ be N sets such that for each $j \in \{1, \dots, N\}$, we have $\int_{I_j^N} f_c(x) dx = \frac{1}{N}$. Since the pre-change distribution has a generalized pdf f that is a mixed distribution between the continuous pdf f_c and point masses over Θ , we also define the sets $I_{N+1}^N = \{\theta_1\}, \dots, I_{N+H}^N = \{\theta_H\}$. We call each of the sets $I_j^N, j = 1, \dots, N + H$, a bin.

Definition 3.2. A distribution $\mathbb{P}^g$, absolutely continuous w.r.t. $\mathbb{P}^f$, is distinguishable from $\mathbb{P}^f$ w.r.t. N if there exists $j \in \{1, \dots, N + H\}$ such that $\int_{I_j^N} f(x)dx \neq \int_{I_j^N} g(x)dx$ where the integral is interpreted in the sense of Lebesgue [106]. The set $D(\mathbb{P}^f, N)$ is the family of distributions distinguishable from $\mathbb{P}^f$ w.r.t. N .

Any distribution $\mathbb{P}^g \neq \mathbb{P}^f$ is distinguishable from $\mathbb{P}^f$ for N sufficiently large. To see why this is true, we define the functions g_N and f_N as

$$g_N(x) \doteq \int_{I_j^N} g(y) dy, \quad f_N(x) \doteq \int_{I_j^N} f(y) dy, \quad (3.2)$$

where j is the unique integer such that $x \in I_j^N$. If there exists $h \in \{0, \dots, H\}$ such that $p_h \neq q_h$ then $g \in D(\mathbb{P}^f, 1)$ with $I_1^1 = (-\infty, \infty)$. Now suppose that $p_h = q_h$ for all $h \in \{0, \dots, H\}$. Let F_c and G_c be the cumulative distribution function of $\mathbb{P}^{f_c}$ and $\mathbb{P}^{g_c}$, respectively. Since both F_c and G_c are continuous and $F_c \neq G_c$, there exists an interval J such that for any $x \in J$, $F_c(x) \neq G_c(x)$. Then for any $N > 1/\int_J f_c(x)dx$, there exists some x such that $g_N(x) \neq f_N(x)$ because otherwise $G_c(x) = F_c(x)$ for any x that is a boundary point of a set in $\{I_j^N : j = 1, \dots, N\}$.

In Section 3.C, we give an exact characterization of N such that a distribution $\mathbb{P}^g$ is distinguishable from $\mathbb{P}^f$ w.r.t. N . We also provide an example to illustrate how N can be determined if additional moment information is available. For the rest of this chapter, we make the following assumption.

Assumption 3.1. For a known positive integer N and pre-change distribution $\mathbb{P}^f$, the post-change distribution $\mathbb{P}^g$ that belongs to the set $D(\mathbb{P}^f, N)$.

In a typical sequential change detection procedure, at each time t , a test statistic $S(t)$ is computed based on the currently available observations $X_1 = x_1, \dots, X_t = x_t$, and the observer makes the decision that a change has occurred at a stopping time τ , where τ is the first t such that $S(t)$ exceeds a pre-determined threshold b :

$$\tau(b) = \inf\{t : S(t) \geq b\}. \quad (3.3)$$

We evaluate the performance of a change detection scheme using the ARL to false alarm and the worst-case ADD (WADD) using Lorden's definitions [7] as follows:

$$\begin{aligned} \text{ARL}(\tau) &= \mathbb{E}_{\infty} [\tau], \\ \text{WADD}(\tau) &= \sup_{\nu \geq 1} \text{ess sup } \mathbb{E}_{\nu} [(\tau - \nu + 1)^+ \mid X_1, \dots, X_{\nu-1}]. \end{aligned}$$

where $\sup$ is the supremum operator and ess sup is the essential supremum operator. The quickest change detection problem can be formulated as the following minimax problem [7]:

$$\begin{aligned} \min_{\tau} \quad & \text{WADD}(\tau), \\ \text{s. t.} \quad & \text{ARL}(\tau) \geq \gamma, \end{aligned} \tag{3.4}$$

for some given $\gamma > 0$.

3.2 Test statistic based on binning

In this section, we derive a test statistic that can be recursively updated. If the post-change distribution belongs to a finite set of distributions, the GLRT statistic is the maximum of a finite set of CuSum statistics, one corresponding to each possible post-change distribution, and therefore has a recursion. To the best of our knowledge, the only other recursive test known in the literature for the case where the post-change distribution is not completely specified and does not belong to a finite set of distributions is proposed in [109]. The method in [109] requires that both the pre- and post-change distributions belong to a single parameter exponential family.

Suppose that the post-change distribution $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ and we observe the sequence $X_1 = x_1, X_2 = x_2, \dots$. We propose to process the observations using the binning approach instead of using raw data. If g_N and f_N , as defined in (3.2), are

both known, Page's CuSum test statistic [6] for the binned observations is

$$S(t) = \max_{1 \leq k \leq t+1} \sum_{j=k}^t \log \frac{g_N(x_j)}{f_N(x_j)}. \quad (3.5)$$

Note that $S(t)$ in (3.5) takes the value 0 if $k = t+1$ is the maximizer. The test statistic $S(t)$ has a convenient recursion $S(t+1) = \max\{S(t) + \log(g_N(x_{t+1})/f_N(x_{t+1})), 0\}$.

In our problem formulation, g_N is unknown. We thus replace $g_N(x_i)$ in (3.5) with its maximum likelihood estimator when $\nu \leq k$

$$g_N^{k:t}(x_i) = \frac{|\{x_r : k \leq r \leq t \text{ and } x_r \in I_j^N\}|}{t - k + 1},$$

where j is the unique integer such that $x_i \in I_j^N$. Note that in computing $g_N^{k:t}(x_i)$, we use only the samples $x_k, \dots, x_t$. We then have the test statistic

$$\max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{g_N^{k:t}(x_i)}{f_N(x_i)}. \quad (3.6)$$

In the case where $t - k + 1$ is small, the maximum likelihood estimator $g_N^{k:t}$ tends to over-fit the observed data and evaluating the likelihood of x_i using $g_N^{k:t}(x_i)$ biases the test statistic. Furthermore, this couples the estimation of g_N and the instantaneous likelihood ratio $\frac{g_N(x_i)}{f_N(x_i)}$. In order to compensate for this over-fitting, we choose not to include observations $x_i, \dots, x_t$ in the estimation of g_N . This also decouples the estimation of g_N and the likelihood ratio $\frac{g_N(x_i)}{f_N(x_i)}$. However, if x_i is the first observation occurring in the set I_j^N , we have $g_N^{k:i-1}(x_i) = 0$.

To avoid this, we regularize the estimate by adding a regularization count when $k \leq i - 1$ and set $g_N^{k:i-1}$ as f_N when $i - 1 < k$ by defining the regularized version of $g_N^{k:i-1}$ as

$$\widehat{g}_N^{k:i-1}(x_i) = \begin{cases} \frac{|\{x_r : k \leq r \leq i-1 \text{ and } x_r \in I_j^N\}| + R}{(N+H)R + i - k} & \text{if } k \leq i - 1, \\ f_N(x_i) & \text{otherwise,} \end{cases} \quad (3.7)$$

where R is a fixed positive constant, and j is the unique integer such that $x_i \in I_j^N$.

Our test statistic then becomes

$$\widehat{S}_N(t) = \max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{\widehat{g}_N^{k:i-1}(x_i)}{f_N(x_i)}, \quad (3.8)$$

with the stopping time

$$\widehat{\tau}(b) = \inf\{t : \widehat{S}_N(t) > b\}. \quad (3.9)$$

In practice, R is chosen to be of the order of N so that $\widehat{g}_N^{k:i-1}(x)$ approaches $1/N$ as $N \rightarrow \infty$. This controls the variability of (3.8) by controlling the range of values that $\log \frac{\widehat{g}_N^{k:i-1}(x_i)}{f_N(x_i)}$ can take. Computation of the test statistic (3.8) is inefficient as the estimator $\widehat{g}_N^{k:i-1}$ needs to be recomputed each time a new observation $X_t = x_t$ is made, leading to computational complexity increasing linearly w.r.t. t . One way to prevent this increase in computational complexity is by searching for a change point from the previous most likely change point rather than from $t = 1$, and also using observations from the previous most likely change point to the current observation to update the estimator for g . Our proposed BG-CuSum test statistic $\widetilde{S}_N$ and test $\widetilde{\tau}$ are defined as follows: For each $t \geq 1$,

$$\widetilde{S}_N(t) = \max_{\lambda_{t-1} \leq k \leq t+1} \sum_{i=k}^t \log \frac{\widehat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)}, \quad (3.10)$$

$$\lambda_t = \max \left\{ \arg \max_{\substack{\lambda_{t-1} \leq k \leq t+1 \\ k \neq \lambda_{t-1}+1}} \sum_{i=k}^t \log \frac{\widehat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)} \right\}, \quad (3.11)$$

$$\widetilde{\tau}(b) = \inf\{t : \widetilde{S}_N(t) \geq b\}, \quad (3.12)$$

where $\lambda_0 = 1$, and b is a fixed threshold. The outer maximum in (3.11) is to ensure that λ_t is uniquely defined when there is more than one maximizer. Due to the

design of the estimator (3.7), we have $\hat{g}_N^{\lambda_{t-1}:\lambda_{t-1}-1}(x_{\lambda_{t-1}}) = f_N(x_{\lambda_{t-1}})$ and

$$\sum_{i=\lambda_{t-1}}^t \log \frac{\hat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)} = \sum_{i=\lambda_{t-1}+1}^t \log \frac{\hat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)}.$$

Thus, if $k = \lambda_{t-1} + 1$ maximizes the sum $\sum_{i=k}^t \log \frac{\hat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)}$, $k = \lambda_{t-1}$ also maximizes it. For this case, we choose the most likely change-point to be λ_{t-1} rather than $\lambda_{t-1} + 1$, as defined in (3.11). Note also that if $\lambda_{t-1} = t$, then we have $\tilde{S}_N(t) = \tilde{S}_N(t-1) = 0$, i.e., it takes more than one sample for our test statistic to move away from the zero boundary once it hits it. This is due to the way we define our estimator $\hat{g}_N^{\lambda_{t-1}:t-1}$ in (3.7), which at time t utilizes samples starting from the last most likely change-point $\lambda_{t-1} = t$ to the previous time $t-1$ to estimate the post-change distribution, i.e., it simply uses f_N as the estimator. From (3.11), we then have $\lambda_t = t$. Furthermore, if $\tilde{S}_N(t-1) = 0$ while $\tilde{S}_N(t) > 0$, then we must have $\lambda_{t-1} = t-1$. Lastly, it can be shown that $\lambda_t = \lambda_{t-1}$ whenever, $\tilde{S}_N(t) > 0$. Using these properties, the test statistic $\tilde{S}_N(t)$ can be efficiently computed using a recursive update as shown in the following result.

Theorem 3.1. For each $t \geq 0$, we have the update formula

$$\tilde{S}_N(t+1) = \max \left\{ \tilde{S}_N(t) + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})}, 0 \right\}, \quad (3.13)$$

$$\lambda_{t+1} = \begin{cases} \lambda_t & \text{if } \tilde{S}_N(t) + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})} > 0 \text{ or } \lambda_t = t+1, \\ t+2 & \text{otherwise,} \end{cases} \quad (3.14)$$

where $\tilde{S}_N(0) = 0$ and $\lambda_0 = 1$.

Proof. See Appendix 3.A. □

Similar to the CuSum test, the renewal property of the test statistic and the fact that it is non-negative implies that the worst case change-point ν for the ADD is at $\nu = 0$. We compare the performance of the stopping time (3.9) with that of (3.12) using simulations in Section 3.4.

3.3 Properties of the BG-CuSum statistic

In this section, we present some properties of the BG-CuSum statistic in order to give insights into the asymptotic behavior of the ARL and ADD of our test. The proofs for the results in this section are provided in Appendix 3.B.

3.3.1 Estimating a lower bound for the ARL

In applications, a practitioner is required to set a threshold b for the problem of interest. Thus, it is of practical interest to have an estimate of the ARL of our test w.r.t. the threshold b . This is even more important in our context as it is not possible to set the threshold b w.r.t. the WADD since it varies with the unknown post-change distribution. In this subsection, we derive a lower bound for the ARL of the BG-CuSum test.

Define the stopping time

$$\zeta(b) = \inf\{t : \tilde{S}_N(t) \geq b \text{ or } (\tilde{S}_N(t) \leq 0 \text{ and } t > 2)\}.$$

Note that the condition $t > 2$ is used due to the lag in our test statistic; see the discussion after (3.12). We have the following lower bound for the ARL of the BG-CuSum test.

Proposition 3.1. For any threshold $b > 0$, we have

$$\text{ARL}(\tilde{\tau}(b)) = \frac{\mathbb{E}_\infty [\zeta(b)]}{\mathbb{P}_\infty \left(\tilde{S}_N(\zeta(b)) \geq b \right)} \geq e^b. \quad (3.15)$$

3.3.2 Growth rate of the BG-CuSum statistic

We study the error bounds of the growth rate of the BG-CuSum statistic $\tilde{S}_N$ under the assumption that the observed samples $X_1 = x_1, X_2 = x_2, \dots$ are generated by the post-change distribution $\mathbb{P}^g$. We show that the growth rate $\tilde{S}_N(t)/t$ converges

to the Kullback-Leibler divergence $D(g_N \| f_N)$ r -quickly (see Section 2.4.3 of [16]) for $r = 1$ under $\mathbb{P}_1$, which implies almost sure convergence. We then provide some heuristic insights into the asymptotic trade-off between the ARL and ADD of our BG-CuSum test as the threshold $b \rightarrow \infty$ in (3.12).

We first show that the regularized sample mean $\hat{g}_N^{1:i}$ in (3.7) is close to g_N with high probability when the sample size is large.

Proposition 3.2. For any $\epsilon \in (0, 1)$ and $x \in \mathbb{R}$, there exists $t_1 \in \mathbb{N}$ such that for all $i \geq t_1$, we have

$$\mathbb{P}_1 \left(|\hat{g}_N^{1:i}(x) - g_N(x)| \geq \epsilon \right) \leq 2e^{-i\epsilon^2/2}.$$

We next show that the instantaneous log-likelihood ratio $\log \frac{\hat{g}_N^{1:i}(x)}{f_N(x)}$ is close to the true log-likelihood ratio $\log \frac{g_N(x)}{f_N(x)}$, in the sense that the probability of a deviation of ϵ decreases to zero exponentially as the number of samples increases.

Proposition 3.3. For any $\epsilon \in (0, 1)$, there exists a $t_2 \in \mathbb{N}$ such that for all $i \geq t_2$, we have

$$\mathbb{P}_1 \left(\left| \log \frac{\hat{g}_N^{1:i}(X_{i+1})}{f_N(X_{i+1})} - \log \frac{g_N(X_{i+1})}{f_N(X_{i+1})} \right| \geq \epsilon \right) \leq 2e^{-i(g_N^{\min}\epsilon)^2/8}$$

where $g_N^{\min} = \min_x g_N(x)$.

Putting Propositions 3.2 and 3.3 together, we obtain the following theorem.

Proposition 3.4. The empirical average $\frac{1}{t} \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)}$ converges to $D(g_N \| f_N)$ r -quickly for $r = 1$ under the distribution $\mathbb{P}_1$ as $t \rightarrow \infty$.

From Proposition 3.4, the probability of the growth rate of the BG-CuSum statistic $\tilde{S}_N(t)/t$ deviating from $D(g_N \| f_N)$ by more than ϵ can be made arbitrarily small by increasing the number of samples t after the change point. Heuristically, this means that the ADD increases linearly at a rate of $1/D(g_N \| f_N)$ w.r.t. the threshold b . Unfortunately, due to technical difficulties introduced by having to estimate

the pmf of the discretized post-change distribution using $\hat{g}_N^{1:i-1}$, we are unable to quantify the asymptotic trade-off between ARL and ADD for our BG-CuSum test $\tilde{\tau}$. In the following, we consider the asymptotic trade-off of the test $\hat{\tau}$ in (3.9) as an approximation of $\tilde{\tau}$, and use simulation results in Section 3.4.3 to verify that a similar trade-off applies for the BG-CuSum test.

Proposition 3.5. The stopping time $\hat{\tau}(b)$ defined in (3.9) satisfies

$$\text{ARL}(\hat{\tau}(b)) \geq e^b,$$

and

$$\text{WADD}(\hat{\tau}(b)) \leq \frac{b}{D(g_N \| f_N)} + o(b) \quad \text{as } b \rightarrow \infty.$$

Furthermore if both $\mathbb{P}^f$ and $\mathbb{P}^g$ are discrete distributions (i.e., $p_0 = q_0 = 0$), the stopping time $\hat{\tau}(b)$ is asymptotically optimal.

3.4 Numerical Results

In this section, we first compare the performance of our proposed BG-CuSum test with two other non-parametric change detection methods in the literature. We first perform simulations, and then we verify the performance of our method on real activity tracking data from [36].

3.4.1 Synthetic data

In our first set of simulations, we set the parameters $H = 0$, $N = 16$, $R = 16$ and choose the pre-change distribution to be the standard normal distribution with pdf $\mathcal{N}(0, 1)$. All the post-change distributions we use in our simulations are in $D(\mathbb{P}^f, N)$ for this choice of parameters.

In our first experiment, we compare the performance of our method with [42], in which it is assumed that the change is an unknown shift in the location parameter.

We also compare the performance of our method with [48], in which it is assumed that the means of the pre- and post-change distributions are different, and both distributions are unknown. In our simulation, we let the post-change distribution be normal distributions with pdf $\mathcal{N}(\delta, 1)$. We control the ARL at 500 and set the change-point $\nu = 300$ for both methods while varying δ . The average detection delay is computed from 50,000 Monte Carlo trials and shown in Table 3.1, where the smallest ADDs for each ARL are highlighted in boldface. We see that our method, despite not assuming that the change is a mean shift, achieves a comparable ADD as the method in [42]. Furthermore, we see that our method outperforms the method in [48].

TABLE 3.1: ADD for post-change distribution with pdf $\mathcal{N}(\delta, 1)$ with ARL = 500.

δ	0.125	0.75	1.5	2.25	3
Hawkins [42]	428.60	18.1	6.6	4.5	3.9
Darkhovskii [48] ($N = 280, b = 0.25, c = 0.1$)	458.71	26.15	13.02	8.55	6.58
BG-CuSum	344.78	17.9	6.6	3.2	2.3

We next consider a shift in variance for the post-change distribution. The method in [42], for example, will not be able to detect this change accurately as it assumes that the change in distribution is a shift in mean. Therefore, we also compare our method with the KS-CPM method [50], which is a non-parametric test that makes use of the Kolmogorov-Smirnov statistic to construct a sequential 2-sample test to test for a change-point. We control the ARL at 500 and change-point $\nu = 300$ for all methods while varying δ for the post-change distribution with pdf $g = \mathcal{N}(0, \delta^2)$. The average detection delays computed from 50,000 Monte Carlo trials are shown in Table 3.2. We see that our method outperforms both [42] and [50] in the ADD. We note that $D(g_N \| f_N)$ decreases as δ varies from 0.2 to 1, and increases as δ varies from 1 to 2. Thus, the observation that the ADD increases as δ varies from 0.2 to 1, and decreases as δ varies from 1 to 2 agrees with our expectation.

Next, we test our method with the Laplace post-change distribution with the pdf $g(x) = \frac{1}{2 \times 0.7071} e^{-\frac{|x|}{0.7071}}$. The location and scale parameter are chosen such that the

TABLE 3.2: ADD for post-change distribution with pdf $\mathcal{N}(0, \delta^2)$ with ARL = 500.

δ	0.2	0.33	0.5	1.5	2
Hawkins [42]	361.3	391.5	438.5	149.6	75.3
KS-CPM [50]	27.2	37.2	84.7	140.6	49.2
BG-CuSum	10.5	17.4	33.3	45.2	21.5

FIG. 3.1: Examples of test statistics $\hat{S}(t)$, $\tilde{S}(t)$ and $\hat{S}(t) - \tilde{S}(t)$ as a function of t for $D(g_N \| f_N) = 2$.

first and second order moments of $\mathbb{P}^g$ and $\mathbb{P}^f$ are equal. We set the ARL = 500 and two different values for the change-point ν . We performed 50,000 Monte Carlo trials to obtain Table 3.3. The results show the method is able to identify the change from a normal to a Laplace distribution with the smallest ADD out of the three methods studied.

TABLE 3.3: ADD for Laplace post-change distribution with ARL = 500.

ν	50	300
Hawkins [42]	592	828
KS-CPM [50]	284	217
BG-CuSum	156	154

FIG. 3.2: Examples of test statistics $\hat{S}(t)$, $\tilde{S}(t)$ and $\hat{S}(t) - \tilde{S}(t)$ as a function of t when the change is in the discrete component such that the pre-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.25\delta_{-1} + 0.25\delta_1$ and post-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.33\delta_{-1} + 0.17\delta_1$.

In Fig. 3.1 and Fig. 3.2, we show how the BG-CuSum test statistic $\tilde{S}_N(t)$ behaves for two different post-change distributions. In Fig. 3.1, the pre-change distribution has pdf $\mathcal{N}(0, 1)$ and the post-change distribution has pdf $\mathcal{N}(0.2, 1)$. In Fig. 3.2, we consider distributions with both discrete and absolutely continuous components. The change is in the discrete component where the pre-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.25\delta_{-1} + 0.25\delta_1$ and the post-change distribution has generalized pdf $0.5\mathcal{N}(0, 1) + 0.33\delta_{-1} + 0.17\delta_1$. We observe, in both cases, that $\tilde{S}_N(t)$ remains low during the pre-change regime and quickly rises in the post-change regime in both cases. Furthermore, the proposed recursive test statistics $\tilde{S}_N$ is observed to track the test statistic $\hat{S}_N$, which has known asymptotic properties as seen in Proposition 3.5.

FIG. 3.3: Graph of ADD against log(ARL) for varying values of N with $f = \mathcal{N}(0, 1)$ and $g = 0.6\mathcal{N}(1, 1) + 0.4\mathcal{N}(-1, 1)$.

FIG. 3.4: Graph of ADD against log(ARL) for varying values of N with $f = \mathcal{N}(0, 1)$ and $g = \mathcal{N}(1, 1)$.

3.4.2 Choice of N

Before applying the BG-CuSum test, the user has to choose an appropriate number of bins N . In Section 3.C, we have provided a procedure for determining a suitable choice of N . In this subsection, we present several results to illustrate the guiding principles for choosing N . First, we compare the ARL-ADD performance of BG-CuSum using different values of N . We performed the experiments with the pre-change distribution with pdf $\mathcal{N}(0, 1)$, and Gaussian mixture model with pdf $0.6\mathcal{N}(1, 1) + 0.4\mathcal{N}(-1, 1)$ as the post-change distribution for $N = 2, 4, 8, 16, 32, 64$, using 5000 Monte Carlo trials to obtain the ADD and ARL. The KL divergence $D(g_N \parallel f_N)$ for respective values of N are 0.0094, 0.730, 0.1164, 0.1420, 0.1565 and 0.1645. From Fig. 3.3, we observe that the performance of BG-CuSum improves as N increases from 2 to 8 and degrades as N increases from 8 to 64. One reason for this is that the benefit of having a larger $D(g_N \parallel f_N)$ does not out-weigh the larger number of samples required to accurately estimate the unknown post-change distribution with pdf g_N when N is large.

3.4.3 Asymptotic behaviour of the ARL and ADD of $\tilde{\tau}$

In Section 3.3, we showed that the growth of the BG-CuSum test statistic can be made arbitrarily close to $D(g_N \parallel f_N)$ with a sufficiently large number of samples t . Heuristically, from Proposition 3.5, using $\hat{S}(t)$ in (3.8) as an approximation of the BG-CuSum test statistic $\tilde{S}_N(t)$, this implies that the average detection delay would grow at a rate of $1/D(g_N \parallel f_N)$. To demonstrate this, we let $N = 64$, $\mathcal{N}(0, 1)$ as the pdf of the pre-change distribution, and g be the pdf of one of several normal distributions with different means and variance 1 so that we have $D(g_N \parallel f_N) = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$, respectively. Therefore we expect the asymptotic gradient of the ADD w.r.t. b to be 1, 2, 3, 4 and 5 respectively. We performed 1000 Monte Carlo trials to estimate the ADD for different values of b . Fig. 3.5 shows the plot of ADD against

FIG. 3.5: Plot of $\Delta(b)$ against threshold b .

b and

$$\Delta(b) = \frac{\text{ADD}(\tau(b+h)) - \text{ADD}(\tau(b))}{h},$$

which approximates the gradient of WADD w.r.t. b . We used a step-size of $h = 12$ to generate Fig. 3.5. We see that $\Delta(b)$ tends to 1, 2, 3, 4 and 5 respectively as b tends to infinity, which agrees with those predicted by our heuristic.

Next, we compare the asymptotic performance of $\tilde{\tau}$ in (3.12) and $\hat{\tau}$ in (3.9). We perform simulations with $N = 4$ and $f = \mathcal{N}(0, 1)$ and g to be the pdf of one of several normal distributions with different means and variance 1 so that we have $D(g_N \| f_N) = 1, 2$ and 3. We perform 5000 Monte Carlo trials to estimate the ARL and ADD of both $\hat{\tau}$ and $\tilde{\tau}$. Fig. 3.6 shows that the ADD of $\tilde{\tau}$ remains close to $\hat{\tau}$ as the ARL becomes large.

FIG. 3.6: Comparison of the trade-off between ADD and $\log(\text{ARL})$ for $\tilde{\tau}$ and $\hat{\tau}$.

3.4.4 WISDM Actitracker Dataset

We now apply our BG-CuSum test on real activity tracking data using the WISDM Actitracker dataset [36], which contains 1,098,207 samples from the accelerometers of Android phones. The dataset was collected from 29 volunteer subjects who each carried an Android phone in their front leg pocket while performing a set of activities such as walking, jogging, ascending stairs, descending stairs, sitting and standing. The signals from the accelerometers were collected at 20 samples per second. The movements generated when each activity by each individual can be assumed to be consistent across time. Hence we assume that the samples generated are i.i.d. for each activity.

To test the effectiveness of the BG-CuSum test on real data, we apply it to detect the changes in the subject's activity. We aim to detect the change from walking activity to any other activity. There are 45 segments in the dataset in which there is a switch from walking to other activities. For each of these segments, we use the

FIG. 3.7: Examples of trial performed for pre-change activity of walking, and post-change activity of (a) jogging and (b) ascending upstairs. The black dotted line indicates the boundary between the pre-change and post-change regimes.

first half of the samples from the walking activity period to learn the pre-change pdf f for the data.

As each sample at time t from the phone's accelerometer is a 3-dimensional vector $\mathbf{a}(t)$, we perform change detection on the sequence $x_t = \|\mathbf{a}(t)\|_2$ instead. We set the number of bins in the BG-CuSum test to be $H = 0$ and $N = 32$. Using the first T samples from each walking activity segment, we estimate the boundaries of the intervals I_j^N , such that $\int_{I_j^N} f(x) dx = 1/N$ by setting $I_1^N = (-\infty, x_{(\lfloor T/N \rfloor)}]$, $I_j^N = (x_{(\lfloor (j-1)T/N \rfloor)}, x_{(\lfloor jT/N \rfloor)}]$ for $1 < j < N - 1$, and $I_N^N = (x_{(\lfloor (N-1)T/N \rfloor)}, \infty)$, where $x_{(n)}$ is the n -th order statistic of x_t .

In order to control the ARL of the BG-CuSum test to be 6000, we set the threshold

FIG. 3.8: Box-plot of the detection delays of each algorithm when the ARL is 6000.

b to be 2.69. The ADD for BG-CuSum on the WISDM Actitracker dataset is found to be about 30.2.

In Fig. 3.7, we present some examples of the performance of BG-CuSum. Similar to the simulations, we observe that in both cases, the test statistic $\tilde{S}_N(t)$ remains low during the pre-change regime and quickly rises in the post-change regime. In Fig. 3.8, we present the notched boxplots of the detection delays of different algorithms when the ARL is set at 6000. We observe that the BG-CuSum test out-performs both the KS-CPM [50] and Hawkins [42] algorithm.

Appendix 3.A Proof of Theorem 3.1

In order to show that $\tilde{S}_N$ can be computed recursively, we require the following lemmas.

Lemma 3.A.1. Suppose $\lambda_p = \lambda_{p+1} = p + 1$ and $\tilde{S}_N(p + 2), \dots, \tilde{S}_N(p + n) > 0$ for some $n \geq 1$, then we have

$$\lambda_{p+2} = \lambda_{p+3} = \dots = \lambda_{p+n} = p + 1.$$

Proof. We prove by contradiction. Suppose that there exists $t \in \{2, \dots, n\}$ such that $\lambda_{p+t} \neq p+1$. Let t_0 be the smallest of all such indices t . Since $\lambda_{p+t_0} \neq p+1$, following the definition of λ_{p+t_0} , we obtain $\lambda_{p+t_0} > \lambda_{p+t_0-1} = p+1$. Furthermore, by (3.11), $\lambda_{p+t_0} \neq \lambda_{p+t_0-1} + 1 = p+2$. Thus we have $\lambda_{p+t_0} > p+2$. From (3.10), we have

$$\begin{aligned} \tilde{S}_N(p+t_0) &= \max_{\lambda_{p+t_0-1} \leq k \leq p+t_0+1} \sum_{i=k}^{p+t_0} \log \frac{\hat{g}_N^{\lambda_{p+t_0-1}:i-1}(x_i)}{f_N(x_i)} \\ &= \max_{p+1 \leq k \leq p+t_0+1} \sum_{i=k}^{p+t_0} \log \frac{\hat{g}_N^{p+1:i-1}(x_i)}{f_N(x_i)} \\ &= \sum_{i=\lambda_{p+t_0}}^{p+t_0} \log \frac{\hat{g}_N^{p+1:i-1}(x_i)}{f_N(x_i)}, \end{aligned} \quad (3.16)$$

where the last equality follows from the definition (3.11). We then obtain

$$\sum_{i=p+1}^{\lambda_{p+t_0}-1} \log \frac{\hat{g}_N^{p+1:i-1}(x_i)}{f_N(x_i)} = \sum_{i=p+1}^{p+t_0} \log \frac{\hat{g}_N^{p+1:i-1}(x_i)}{f_N(x_i)} - \tilde{S}_N(p+t_0) \leq 0, \quad (3.17)$$

where the inequality follows from (3.16). Since $p+1 < \lambda_{p+t_0} - 1 \leq p+t_0$, let $\lambda_{p+t_0} - 1 = p+t_1$ where $2 \leq t_1 \leq t_0 \leq n$. Since t_0 is the smallest t such that $\lambda_{p+t} \neq p+1$, we have $\lambda_{p+t_1-1} = p+1$ and

$$\begin{aligned} \tilde{S}_N(p+t_1) &= \sum_{i=\lambda_{p+t_1}}^{p+t_1} \log \frac{\hat{g}_N^{\lambda_{p+t_1-1}:i-1}(x_i)}{f_N(x_i)} \\ &= \sum_{i=\lambda_{p+t_1}}^{\lambda_{p+t_0}-1} \log \frac{\hat{g}_N^{p+1:i-1}(x_i)}{f_N(x_i)} \leq 0, \end{aligned} \quad (3.18)$$

where the last inequality follows trivially if $t_1 = t_0$ and from (3.17) if $t_1 < t_0$. This inequality contradicts our assumption that $\tilde{S}_N(p+2), \dots, \tilde{S}_N(p+n) > 0$. The lemma is now proved. $\square$

Lemma 3.A.2. If $\tilde{S}_N(t+1) > 0$, then $\tilde{S}_N(t+1) = \tilde{S}_N(t) + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})}$

Proof. We first consider the case where $\tilde{S}_N(t) = 0$. Since $\tilde{S}_N(t+1) > 0$, we must have $\lambda_t = t$. We then obtain

$$\begin{aligned}\tilde{S}_N(t+1) &= \max_{\lambda_t \leq k \leq t+2} \sum_{i=k}^{t+1} \log \frac{\hat{g}_N^{\lambda_t:i-1}(x_i)}{f_N(x_i)} \\ &= \max_{t \leq k \leq t+2} \sum_{i=k}^{t+1} \log \frac{\hat{g}_N^{t:i-1}(x_i)}{f_N(x_i)} \\ &= \log \frac{\hat{g}_N^{t:t}(x_{t+1})}{f_N(x_{t+1})} \\ &= \tilde{S}_N(t) + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})}.\end{aligned}$$

For the case where $\tilde{S}_N(t) > 0$, let $n \geq 0$ be the largest integer such that $\tilde{S}_N(t-n), \dots, \tilde{S}_N(t) > 0$. If $n = 0$, we trivially have $\lambda_{t-1} = \lambda_t = t-1$. If $n > 0$, since $\tilde{S}_N(t-n-1) = 0$, we have $\lambda_{t-n-1} = t-n-1$ and Lemma 3.A.1 yields $\lambda_{t-n} = \dots = \lambda_t = t-n-1$. We then obtain

$$\begin{aligned}\tilde{S}_N(t+1) &= \max_{\lambda_t \leq k \leq t+1} \sum_{i=k}^{t+1} \log \frac{\hat{g}_N^{\lambda_t:i-1}(x_i)}{f_N(x_i)} \\ &= \max_{\lambda_t \leq k \leq t+1} \left\{ \sum_{i=k}^t \log \frac{\hat{g}_N^{\lambda_t:i-1}(x_i)}{f_N(x_i)} + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})} \right\} \\ &= \max_{\lambda_{t-1} \leq k \leq t+1} \left\{ \sum_{i=k}^t \log \frac{\hat{g}_N^{\lambda_{t-1}:i-1}(x_i)}{f_N(x_i)} \right\} + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})} \\ &= \tilde{S}_N(t) + \log \frac{\hat{g}_N^{\lambda_t:t}(x_{t+1})}{f_N(x_{t+1})},\end{aligned}$$

and the proof is complete. $\square$

We are now ready to show that the statistic $\tilde{S}_N$ can be updated recursively. From Lemma 3.A.2 and noting that $\tilde{S}_N(t) \geq 0$ for all t , (3.13) follows immediately. If $\tilde{S}_N(t+1) = 0$ and $\lambda_t \neq t+1$, then $\lambda_{t+1} = t+2$ from (3.11). Next, if $\tilde{S}_N(t+1) = 0$ and $\lambda_t = t+1$, then $\lambda_{t+1} = t+1 = \lambda_t$ from (3.11). Finally, if $\tilde{S}_N(t+1) > 0$, then by Lemma 3.A.1, we have $\lambda_{t+1} = \lambda_t$, and (3.14) follows. The proof is now complete.

Appendix 3.B Proofs of Results in Section 3.3

In this appendix, we let $M = N + H$ and Y_k^j for $k = 1, 2, \dots$ to be i.i.d. random variables such that

$$Y_k^j = \begin{cases} 1 & \text{if } X_k \in I_j^N, \\ 0 & \text{if } X_k \notin I_j^N. \end{cases}$$

Recalling (3.6), we then have

$$g_N(x) = \mathbb{E} [Y_1^j]$$

and the regularized sample mean $\widehat{g}_N^{1:i}(x)$ defined in (3.7) can be written as

$$\widehat{g}_N^{1:i}(x) = \frac{\sum_{k=1}^i Y_k^j + R}{i + MR}$$

where j is the unique integer such that $x \in I_j^N$. We first study the error bounds for the regularized sample mean $\widehat{g}_N^{1:i}(x)$. We then derive error bounds for the instantaneous sample log-likelihood ratio $\log \frac{\widehat{g}_N^{1:i}(x)}{f_N(x)}$. Finally we combine all the results together to derive an error bound for the growth rate of the BG-CuSum statistic.

3.B.1 Proof of Proposition 3.1

Following arguments identical to those in Chapter 2 of [110], it can be shown that

$$\text{ARL}(\widetilde{\tau}(b)) = \frac{\mathbb{E}_\infty [\zeta(b)]}{\mathbb{P}_\infty \left(\widetilde{S}_N(\zeta(b)) \geq b \right)}. \quad (3.19)$$

In order to obtain a lower bound for the ARL, we first note that for any $b > 0$, we have

$$\mathbb{E}_\infty [\zeta(b)] \geq \mathbb{E}_\infty [\zeta(0)] = 1. \quad (3.20)$$

We also have

$$\mathbb{P}_\infty \left(\tilde{S}_N(\zeta(b)) \geq b \right) = \mathbb{P}_\infty \left(\sum_{i=1}^{\zeta(b)} \log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} \geq b \right) \quad (3.21)$$

$$\begin{aligned} &= \mathbb{P}_\infty \left(\prod_{i=1}^{\zeta(b)} \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} \geq e^b \right) \\ &= \sum_{t=1}^{\infty} \mathbb{P}_\infty(E_t), \end{aligned} \quad (3.22)$$

where E_t is the event $\left\{ \prod_{i=1}^t \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} \geq e^b, \zeta(b) = t \right\}$. The equality in (3.21) follows from Theorem 3.1 and noting that since $\zeta(b)$ is the first time $\tilde{S}_N(t)$ exceeds b or falls below 0, $\lambda_i = \lambda_0 = 1$ for all $i = 1, \dots, \zeta(b)$ under the event $\tilde{S}_N(\zeta(b)) \geq b$.

Let $J_i \in \{1, \dots, N+H\}$ denote the index of the bin in $\{I_j^N\}_{j=1}^{N+H}$ that X_i falls into. Let F_t be the event such that $\{J_i\}_{i=1}^t \in F_t$ if and only if $\{X_i\}_{i=1}^t \in E_t$. For any $i \geq 1$ and any sequence $(x_1, \dots, x_i)$ with corresponding bin indices $(j_1, \dots, j_i)$, let $q(j_i \mid j_{1:i-1}) = \hat{g}_N^{1:i-1}(x_i)$. From the Kolmogorov Extension Theorem [111], there exists a probability measure $\mathbb{Q}$ on $\{1, \dots, N+H\}^\mathbb{N}$ such that $\mathbb{Q}(J_1 = j_1, \dots, J_t = j_t) = \prod_{i=1}^t q(j_i \mid j_{1:i-1})$ for all $t \geq 1$. We then have

$$\begin{aligned} \mathbb{P}_\infty(E_t) &= \mathbb{E}_\infty[\mathbf{1}_{E_t}] \\ &\leq \mathbb{E}_\infty \left[e^{-b} \prod_{i=1}^t \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} \mathbf{1}_{E_t} \right] \\ &\leq e^{-b} \sum_{\{j_i\} \in F_t} \prod_{i=1}^t q(j_i \mid j_{1:i-1}) \\ &= e^{-b} \mathbb{Q}(F_t). \end{aligned}$$

From (3.22), we obtain

$$\begin{aligned} \mathbb{P}_\infty \left(\tilde{S}_N(\zeta(b)) \geq b \right) &\leq e^{-b} \sum_{t=1}^{\infty} \mathbb{Q}(F_t) \\ &\leq e^{-b}. \end{aligned} \quad (3.23)$$

where the final inequality follows because $\{F_t\}_{t \in \mathbb{N}}$ are mutually exclusive events. Thus, from (3.19), (3.20) and (3.23), we have

$$\text{ARL}(\tilde{\tau}(b)) = \frac{\mathbb{E}_\infty [\zeta(b)]}{\mathbb{P}_\infty \left(\tilde{S}_N(\zeta(b)) \geq b \right)} \geq e^b. \quad (3.24)$$

The proof is now complete.

3.B.2 Proof of Proposition 3.2

In the design of the BG-CuSum test statistic, we replace the maximum likelihood estimator $g_N^{1:i}$ with its regularized sample mean $\hat{g}_N^{1:i}$. In this subsection, we derive a bound for the probability that $\hat{g}_N^{1:i}(x)$ deviates from $g_N(x)$ by at least $\epsilon > 0$. We start with a few elementary lemmas.

Lemma 3.B.1. For any $\epsilon \geq 0$, $N \geq 1$, and $x \in \mathbb{R}$, we have

$$\mathbb{P}_1 \left(\left| \hat{g}_N^{1:i}(x) - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} \right| \geq \epsilon \right) \leq 2e^{-2i\epsilon^2}$$

where j is the unique integer such that $x \in I_j^N$.

Proof. By applying Hoeffding's inequality [112], we obtain

$$\begin{aligned} & \mathbb{P}_1 \left(\left| \hat{g}_N^{1:i}(x) - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} \right| \geq \epsilon \right) \\ &= \mathbb{P}_1 \left(\left| \frac{\sum_{k=1}^i Y_k^j + R}{i + MR} - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} \right| \geq \epsilon \right) \\ &= \mathbb{P}_1 \left(\left| \frac{\sum_{k=1}^i Y_k^j + R}{i} - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i} \right| \geq \frac{i + MR}{i} \epsilon \right) \\ &\leq 2e^{-2i \left(\frac{i + MR}{i} \right)^2 \epsilon^2} \leq 2e^{-2i\epsilon^2}, \end{aligned}$$

and the proof is complete. $\square$

Lemma 3.B.2. For any $\epsilon \in (0, 1)$ and $x \in \mathbb{R}$, there exists $i_0 \in \mathbb{N}$ such that for all $i \geq i_0$, we have

$$\left| \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} - g_N(x) \right| < \epsilon$$

where j is the unique integer such that $x \in I_j^N$.

Proof. For $i \geq (1 - \epsilon)MR/\epsilon$, we have

$$\begin{aligned} \left| \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} - g_N(x) \right| &= \left| \frac{ig_N(x) + R}{i + MR} - g_N(x) \right| \\ &= \left| \frac{R - MRg_N(x)}{i + MR} \right| \\ &= \frac{R}{i + MR} |1 - Mg_N(x)| \\ &\leq \frac{MR}{i + MR} \\ &\leq \epsilon, \end{aligned}$$

and the proof is complete. $\square$

Putting everything together, we now proceed to the proof of Proposition 3.2. Given $0 < \epsilon < 1$, by Lemma 3.B.2 there exists $i_1 > 0$ such that for all $i \geq i_1$, we have

$$\left| \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} - g_N(x) \right| < \frac{\epsilon}{2}.$$

Therefore, we obtain

$$\begin{aligned} &\mathbb{P}_1 \left(\left| \hat{g}_N^{1:i}(x) - g_N(x) \right| \geq \epsilon \right) \\ &= \mathbb{P}_1 \left(\left| \hat{g}_N^{1:i}(x) - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} + \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} - g_N(x) \right| \geq \epsilon \right) \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{P}_1 \left(\left| \widehat{g}_N^{1:t}(x) - \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} \right| + \left| \frac{\mathbb{E} \left[\sum_{k=1}^i Y_k^j \right] + R}{i + MR} - g_N(x) \right| \geq \epsilon \right) \\
&\leq \mathbb{P}_1 \left(\left| \widehat{g}_N^{1:i}(x) - \frac{\mathbb{E} \left[\sum_{k=1}^t Y_k^j \right] + R}{i + MR} \right| \geq \frac{\epsilon}{2} \right) \\
&\leq 2e^{-\frac{1}{2}i\epsilon^2},
\end{aligned}$$

where the last inequality follows from Lemma 3.B.1, and the proposition is proved.

3.B.3 Proof of Proposition 3.3

In this subsection, we use previous results on the regularized sample mean $\widehat{g}_N^{1:i}$ to study the instantaneous log-likelihood ratio $\log \frac{\widehat{g}_N^{1:i}(x)}{f_N(x)}$ used in the BG-CuSum test statistic.

Lemma 3.B.3. For any $\epsilon \in (0, 1)$, there exists a $t_2 \in \mathbb{N}$ such that for all $i \geq t_2$ and any $x \in \mathbb{R}$, we have

$$\mathbb{P}_1 \left(\left| \log \widehat{g}_N^{1:i}(x) - \log g_N(x) \right| \geq \epsilon \right) \leq 2e^{-i(g_N^{\min} \epsilon)^2/8}$$

where $g_N^{\min} = \min_x g_N(x)$.

Proof. Using the inequality $\log x \leq x - 1$, we obtain for each $x \in \mathbb{R}$,

$$\mathbb{P}_1 \left(\log \frac{\widehat{g}_N^{1:i}(x)}{g_N(x)} \geq \epsilon \right) \leq \mathbb{P} \left(\widehat{g}_N^{1:i}(x) - g_N(x) \geq \epsilon g_N(x) \right)$$

and

$$\mathbb{P}_1 \left(\log \frac{g_N(x)}{\widehat{g}_N^{1:i}(x)} \geq \epsilon \right) \leq \mathbb{P} \left(g_N(x) - \widehat{g}_N^{1:i}(x) \geq \frac{\epsilon}{1 + \epsilon} g_N(x) \right).$$

Therefore, we have

$$\begin{aligned}
& \mathbb{P}_1 \left(|\log \widehat{g}_N^{1:i}(x) - \log g_N(x)| \geq \epsilon \right) \\
&= \mathbb{P}_1 \left(\log \frac{\widehat{g}_N^{1:i}(x)}{g_N(x)} \geq \epsilon \right) + \mathbb{P}_1 \left(\log \frac{g_N(x)}{\widehat{g}_N^{1:i}(x)} \geq \epsilon \right) \\
&\leq \mathbb{P}_1 \left(\widehat{g}_N^{1:i}(x) - g_N(x) \geq \epsilon g_N(x) \right) \\
&\quad + \mathbb{P}_1 \left(g_N(x) - \widehat{g}_N^{1:i}(x) \geq \frac{\epsilon}{1+\epsilon} g_N(x) \right) \\
&\leq \mathbb{P}_1 \left(|\widehat{g}_N^{1:i}(x) - g_N(x)| \geq \min \left(\epsilon g_N(x), \frac{\epsilon}{1+\epsilon} g_N(x) \right) \right) \\
&\leq \mathbb{P}_1 \left(|\widehat{g}_N^{1:i}(x) - g_N(x)| \geq \frac{1}{2} \epsilon g_N(x) \right) \\
&\leq 2e^{-i(g_N(x)\epsilon)^2/8} \leq 2e^{-i(g_N^{\min}\epsilon)^2/8},
\end{aligned}$$

where the last inequality follows from Proposition 3.2 for all $i \geq t_1^x$, where $t_1^j \in \mathbb{N}$ is chosen to be sufficiently large. Taking $t_2 = \max_{x \in \mathbb{R}} t_1^x$, the lemma follows. $\square$

We now proceed to the proof of Proposition 3.3. Using the law of total probability and Lemma 3.B.3, we have for $i \geq t_2$, where t_2 is as given in Lemma 3.B.3,

$$\begin{aligned}
& \mathbb{P}_1 \left(\left| \log \frac{\widehat{g}_N^{1:i}(X_{i+1})}{f_N(X_{i+1})} - \log \frac{g_N(X_{i+1})}{f_N(X_{i+1})} \right| \geq \epsilon \right) \\
&= \sum_{j=1}^M \mathbb{P}_1 \left(|\log \widehat{g}_N^{1:i}(X_{i+1}) - \log g_N(X_{i+1})| \geq \epsilon \mid X_{i+1} \in I_j^N \right) \\
&\quad \times \mathbb{P}_1 \left(X_{i+1} \in I_j^N \right) \\
&= \sum_{j=1}^M \mathbb{P}_1 \left(|\log \widehat{g}_N^{1:i}(X_{i+1}) - \log g_N(X_{i+1})| \geq \epsilon \mid X_{i+1} \in I_j^N \right) \\
&\quad \times \mathbb{P}_1 \left(X_{i+1} \in I_j^N \right) \\
&= \sum_{j=1}^M \mathbb{P}_1 \left(|\log \widehat{g}_N^{1:i}(X_{i+1}) - \log g_N(X_{i+1})| \geq \epsilon \right) \\
&\quad \times \mathbb{P}_1 \left(X_{i+1} \in I_j^N \right) \\
&\leq \sum_{j=1}^M 2e^{-\frac{i}{8}(g_N^{\min}\epsilon)^2} \mathbb{P}_1 \left(X_{i+1} \in I_j^N \right)
\end{aligned}$$

$$= 2e^{-i(g_N^{\min}\epsilon)^2/8},$$

and the proof is complete.

3.B.4 Proof of Proposition 3.4

In this subsection, we use results derived in the previous two subsections to study the growth rate $\tilde{S}_N(t)/t$ of the BG-CuSum test statistic.

Lemma 3.B.4. For any $\epsilon \in (0, 1)$, there exists a $t_3 \in \mathbb{N}$ such that for all $t \geq t_3$, we have

$$\mathbb{P}_1 \left(\left| \frac{1}{t} \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| \geq \epsilon \right) \leq c_2 e^{-c_1 t^{1-\epsilon}},$$

where c_1, c_2 are positive constants.

Proof. Let $l = \lceil t^{1-\epsilon} \rceil$. For all $i \leq l$, we have from (3.7),

$$\hat{g}_N^{1:i-1}(X_i) \geq \frac{R}{l + MR} \quad \mathbb{P}_1\text{-a.s.},$$

which yields

$$|\log \hat{g}_N^{1:i-1}(X_i) - \log g_N(X_i)| \leq \log \frac{l + MR}{R} + |\log g_N^{\min}|, \quad (3.25)$$

where $g_N^{\min} = \min_j g_N(j) > 0$ since $f_N(j) = 1/N$ for all j and is absolutely continuous w.r.t. g_N . There exists $t_3 \in \mathbb{N}$ such that for all $t \geq t_3$,

$$\begin{aligned} & \frac{l}{t} \left(\log \frac{l + MR}{R} + |\log g_N^{\min}| \right) \\ & \leq \frac{t^{1-\epsilon} + 1}{t} \left(\log \frac{t^{1-\epsilon} + 1 + MR}{R} + |\log g_N^{\min}| \right) \\ & \leq \left(t^{-\epsilon} + \frac{1}{t} \right) \left(\log \frac{t^{1-\epsilon} + 1 + MR}{R} + |\log g_N^{\min}| \right) \leq \frac{\epsilon}{2}. \end{aligned}$$

We then obtain for $t \geq t_3$,

$$\begin{aligned}
& \mathbb{P}_1 \left(\left| \frac{1}{t} \sum_{i=1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| \geq \epsilon \right) \\
& \leq \mathbb{P}_1 \left(\left| \frac{1}{t-l} \sum_{i=l+1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| + \frac{l}{t} \left(\log \frac{l+MR}{R} + |\log g_N^{\min}| \right) \geq \epsilon \right) \\
& \leq \mathbb{P}_1 \left(\left| \frac{1}{t-l} \sum_{i=l+1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| \geq \frac{\epsilon}{2} \right) \\
& \leq \sum_{i=l+1}^t \mathbb{P}_1 \left(|\log \widehat{g}_N^{1:i-1}(X_i) - \log g_N(X_i)| \geq \frac{\epsilon}{2} \right) \\
& \leq 2 \sum_{i=l+1}^t e^{-i(g_N^{\min}\epsilon)^2/8} \\
& \leq c_2 e^{-c_1 l} \leq c_2 e^{-c_1 t^{1-\epsilon}},
\end{aligned}$$

where the penultimate inequality follows from Proposition 3.3, $c_1 = (g_N^{\min}\epsilon)^2/8$, and $c_2 = 2(1 - e^{-c_1})^{-1}$. The lemma is now proved. $\square$

Finally, we are ready to prove Proposition 3.4. We begin by showing that $\frac{1}{t} \sum_{i=1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)}$ converges to zero r-quickly under the distribution $\mathbb{P}_1$. For any $\epsilon \in (0, 1)$, let

$$L_\epsilon = \sup \left\{ t : \left| \frac{1}{t} \sum_{i=1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| > \epsilon \right\}.$$

We have

$$\begin{aligned}
\mathbb{E}_1 [L_\epsilon] &= \sum_{n=1}^{\infty} \mathbb{P}_1 (L_\epsilon \geq n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_1 \left(\left| \frac{1}{t} \sum_{i=1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| \geq \epsilon, \text{ for some } t \geq n \right) \\
&\leq \sum_{n=1}^{\infty} \sum_{t=n}^{\infty} \mathbb{P}_1 \left(\left| \frac{1}{t} \sum_{i=1}^t \log \frac{\widehat{g}_N^{1:i-1}(X_i)}{g_N(X_i)} \right| \geq \epsilon \right) \\
&= \sum_{n=1}^{\infty} \sum_{t=n}^{\infty} c_2 e^{-c_1 t^{1-\epsilon}} \\
&= c_2 \sum_{n=1}^{\infty} n e^{-c_1 n^{1-\epsilon}} < \infty,
\end{aligned}$$

where the penultimate equality follows from Lemma 3.B.4, and c_1, c_2 are positive constants. Thus, $\frac{1}{t} \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{g_N(X_i)}$ converges to zero r -quickly for $r = 1$ under the distribution $\mathbb{P}_1$.

Since $\mathbb{E}_1 \left[\left(\log \frac{g_N(X)}{f_N(X)} \right)^2 \right] < \infty$, from Theorem 2.4.4 of [16], $\frac{1}{t} \sum_{i=1}^t \log \frac{g_N(X_i)}{f_N(X_i)}$ converges to $D(g_N \| f_N)$ r -quickly for $r = 1$ under the distribution $\mathbb{P}_1$. Therefore, $\frac{1}{t} \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)}$ converges to $D(g_N \| f_N)$ r -quickly for $r = 1$ under $\mathbb{P}_1$ and the proof is complete.

3.B.5 Proof of Proposition 3.5

Let

$$\hat{T} = \inf \left\{ t : \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} > b \right\}.$$

Using Proposition 3.4 and Corollary 3.4.1 in [16], we obtain

$$\mathbb{E}_1 [\hat{T}] \sim \frac{b}{D(g_N \| f_N)} \quad \text{as } b \rightarrow \infty.$$

Using arguments similar to those that led to Eq (3.23), we obtain

$$\mathbb{P}_\infty (\hat{T} < \infty) = \sum_{t=1}^{\infty} \mathbb{P}_\infty (\hat{T} = t) \leq e^{-b}.$$

Applying results from Theorem 6.16 in [18] to translate our understanding of $\hat{T}$ onto $\hat{\tau}(b)$, we obtain

$$\text{ARL}(\hat{\tau}(b)) \geq \frac{1}{\mathbb{P}_\infty (\hat{T} < \infty)} = e^b$$

and that

$$\text{WADD}(\hat{\tau}(b)) \leq \mathbb{E}_1 [\hat{T}] \sim \frac{b}{D(g_N \| f_N)} \quad \text{as } b \rightarrow \infty.$$

For the case where f and g are discrete distributions, we have $f_N = f$ and $g_N = g$, so that these bounds coincide with the bounds for the CuSum stopping time when

both f and g are known. Thus, for this case, the test is asymptotically optimal and the theorem is now proved.

Appendix 3.C Characterization of the number of bins N

In this section, we give an exact characterization of N such that a distribution $\mathbb{P}^g$ is distinguishable from $\mathbb{P}^f$ w.r.t. N . We also provide an example to illustrate how N can be determined if additional moment information is available.

3.C.1 Properties of $D(\mathbb{P}^f, N)$

In the design of the BG-CuSum test, we assume that the post-change distribution $\mathbb{P}^g$ is distinguishable from $\mathbb{P}^f$ w.r.t. N . In this section, we derive some properties of $D(\mathbb{P}^f, N)$ and give an example on how to choose N for a particular family of post-change distributions.

Let F_c be the cumulative density function (cdf) of the continuous part of the pre-change distribution $\mathbb{P}^f$. Let G_c be the cdf of the continuous part of the unknown post-change distribution $\mathbb{P}^g$. Let I be the image under F_c of the zero set of $F_c - G_c$ defined as:

$$I = \{F_c(x) : F_c(x) - G_c(x) = 0, x \in \mathbb{R} \cup \{-\infty, \infty\}\}, \quad (3.26)$$

where the terms $F_c(\infty), F_c(-\infty), G_c(\infty), G_c(-\infty)$ are defined to be the limits of F_c and G_c as x tends to ∞ or $-\infty$ respectively.

We first begin by deriving a necessary and sufficient condition for $\mathbb{P}^g$ to be distinguishable from $\mathbb{P}^f$ w.r.t. N .

Proposition 3.C.1. The distribution $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ if and only if $p_h \neq q_h$ for some $h \in \{0, \dots, H\}$ or

$$\left\{ \frac{i}{N} : 0 \leq i \leq N \right\} \setminus I \neq \emptyset. \quad (3.27)$$

Proof. Let the intervals $(a_0, a_1], (a_1, a_2], \dots, (a_{N-1}, a_N)$ with $a_0 = -\infty, a_N = \infty$ be such that for each $i \in \{1, \dots, N-1\}$, $F_c(a_i) = \frac{i}{N}$.

If $p_h \neq q_h$ for some $h \in \{0, \dots, H\}$, then $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ trivially. Suppose now that $p_h = q_h$ for all $h \in \{0, \dots, H\}$, and (3.27) holds. There exists a_i such that $F_c(a_i) = i/N$ and $G_c(a_i) \neq i/N$. This implies that $\sum_{j=1}^i f_N(a_j) \neq \sum_{j=1}^i g_N(a_j)$. Therefore, there exists at least one j such that $f_N(a_j) \neq g_N(a_j)$. Hence g is distinguishable from f w.r.t. N .

On the other hand, suppose now that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ and $p_h = q_h$ for all $h \in \{0, \dots, H\}$. Then for some $j \in \{1, \dots, N-1\}$,

$$F_c(a_j) - F_c(a_{j-1}) \neq G_c(a_j) - G_c(a_{j-1}). \quad (3.28)$$

Letting j_0 be the smallest of such j satisfying (3.28), we have $F_c(a_{j_0}) \neq G_c(a_{j_0})$. Therefore, $a_{j_0} \notin I$. Since F_c is injective, $j_0/N \notin I$. Therefore the left hand side of (3.27) is non-empty. The proof is now complete. $\square$

An easy application of Proposition 3.C.1 relates the number of elements of the set I and N for $\mathbb{P}^g$ to be distinguishable from $\mathbb{P}^f$ w.r.t. N .

Corollary 3.C.1. If $|I|$ is finite, then $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ for all $N \geq |I|$.

Proof. For any $N \geq |I|$, counting the number of elements in each set we obtain

$$\left| \left\{ \frac{i}{N} : 0 \leq i \leq N \right\} \setminus I \right| \geq 1.$$

From Proposition 3.C.1, we have $\mathbb{P}^g \in D(\mathbb{P}^f, N)$, and the corollary follows. $\square$

As an example, if the post-change distribution $\mathbb{P}^g$ is $\mathbb{P}^f$ shifted in mean, then $|I| = 1$ and we can choose $N = 1$. In general, N can be chosen based on prior statistical information about $\mathbb{P}^g$. As an illustration, we consider post-change distributions that satisfy the following assumption in the remainder of this section.

Assumption 3.2. For some positive integer k and $\epsilon > 0$, we have $|\mathbb{E}_g[X^k] - \mathbb{E}_f[X^k]| > \epsilon$. Furthermore, there exist $C, \xi > 0$ such that $f_c(x) \leq C|x|^{-k-1-\xi}$ and $g_c(x) \leq C|x|^{-k-1-\xi}$ for all $x \in \mathbb{R}$.

Note that Assumption 3.2 does not require us to know the k -th moment of $\mathbb{P}^g$. The first part of Assumption 3.2 is fairly reasonable as k and ϵ are guaranteed to exist when the moment generating functions for both $\mathbb{P}^f$ and $\mathbb{P}^g$ exist. The second part of Assumption 3.2 will be automatically satisfied when f and g are compactly supported. This is usually the case for sensor readings obtained from digital sensors due to the finite dynamic range of such sensors. Thus, Assumption 3.2 is satisfied for the class for compactly support distributions which are commonly encountered when modeling observations obtained from digital sensors.

In the following, we present a result that allows us to derive an algorithm for selecting N so that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$. It gives a lower bound on $|I|$ if the first $k - 1$ moments of $\mathbb{P}^f$ and $\mathbb{P}^g$ are equal.

Proposition 3.C.2. Suppose $\mathbb{P}^{g_c} \neq \mathbb{P}^{f_c}$, the first $k - 1$ moments of $\mathbb{P}^{f_c}$ and $\mathbb{P}^{g_c}$ are equal. Then, $|I| \geq k$.

Proof. The claim is trivial if $k < 2$. Therefore, we consider only the case where $k \geq 2$, and proceed by contradiction. Suppose $|I| < k$.

We define a sign change to be a compact interval $[a, b]$ (with possibly $b = a$, in which case $[a, b]$ is a singleton set) such that $F_c(x) - G_c(x) = 0$ for all $x \in [a, b]$, and there exist $x_1 < a \leq b < x_2$ such that $(F_c(x_1) - G_c(x_1))(F_c(x_2) - G_c(x_2)) < 0$. Then, from the definition of I in (3.26), since $F_c(x) - G_c(x)$ has at most $|I| - 2 < k - 2$ sign changes, there exists a degree $k - 1$ polynomial $P(x) = \sum_{i=0}^{k-1} a_i x^i$ such that its derivative $P'(x)$ satisfies $P'(x)(F_c(x) - G_c(x)) > 0$ for all $x \in Z = \mathbb{R} \setminus \{x \in \mathbb{R} :$

$F_c(x) = G_c(x)\}$. Note that since $g_c \neq f_c$ according to our definition, the set Z has non-zero Lebesgue measure. Integrating by parts, we have

$$\begin{aligned} 0 &< \int_{-\infty}^{\infty} P'(x) (F_c(x) - G_c(x)) \, dx \\ &= - \int_{-\infty}^{\infty} P(x) (f_c(x) - g_c(x)) \, dx \\ &= - \sum_{i=0}^{k-1} a_i \int_{-\infty}^{\infty} x^i (f_c(x) - g_c(x)) \, dx = 0, \end{aligned}$$

where the last equality follows from the assumption that the first $k - 1$ moments of $\mathbb{P}^{f_c}$ and $\mathbb{P}^{g_c}$ are equal. This gives us a contradiction. Therefore, $|I| \geq k$, and the proof is complete. $\square$

The following theorem gives us a method to search for N such that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$.

Theorem 3.C.1. Suppose Assumption 3.2 holds. If $\mathbb{P}^g \notin D(\mathbb{P}^f, N)$, we have $m_N \leq \mathbb{E}_g [X^k] \leq M_N$, where

$$M_N = p_0 \left(\frac{1}{N} \sum_{i=2}^{N-1} \max_{x \in I_i^N} (x^k) + \int_{I_1^N \cup I_N^N} \max \left(\frac{C x^k}{|x|^{k+1+\xi}}, 0 \right) dx \right) + \sum_{h=1}^H p_h \theta_h^k, \quad (3.29)$$

$$m_N = p_0 \left(\frac{1}{N} \sum_{i=2}^{N-1} \min_{x \in I_i^N} (x^k) \right) + \sum_{h=1}^H p_h \theta_h^k \quad (3.30)$$

Furthermore, $\lim_{d \rightarrow \infty} M_{2^d} = \lim_{d \rightarrow \infty} m_{2^d} = \mathbb{E}_f [X^k]$.

Proof. See Section 3.C.2. $\square$

Theorem 3.C.1 shows that there exists a sufficiently large N so that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$. To determine a suitable N , we note from Corollary 3.C.1 that $\mathbb{P}^g$ is distinguishable from $\mathbb{P}^f$ w.r.t. N for any $N \geq |I|$. From Proposition 3.C.2, a candidate to start the search of N such that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ would be $N = k$. A procedure to find N so that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ is given in Algorithm 1, which is guaranteed to stop after a finite number of iterations due to Theorem 3.C.1.

Algorithm 1 Given Assumption 3.2, compute N so that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$.

Initialize:

Set $N := k$

while ($M_N > \mathbb{E}_f[X^k] + \epsilon$ or $m_N < \mathbb{E}_f[X^k] - \epsilon$) **do**

$N := N + 1$

 Compute the bins I_i^N , $i = 1, \dots, N + H$, according to Definition 3.1.

 Compute M_N and m_N according to (3.29) and (3.30), respectively.

end while

return N

3.C.2 Proof of Theorem 3.C.1

We break the proof into two parts. In Proposition 3.C.3, we derive a upper bound M_N and lower bound m_N on the k -th moment of $\mathbb{P}^g$ if $\mathbb{P}^g \notin D(\mathbb{P}^f, N)$. Then, in Proposition 3.C.4, we show that the sub-sequence M_{2^u} and m_{2^u} converges to $\mathbb{E}_f[X^k]$ as $u \rightarrow \infty$. We let $\mathbf{1}_A(x)$ be the indicator function for the set A .

Proposition 3.C.3. Under the setup of Theorem 3.C.1, we have $m_N \leq \mathbb{E}_g[X^k] \leq M_N$.

Proof. Since $\mathbb{P}^g \notin D(\mathbb{P}^f, N)$, $p_h = q_h$ for $h \in \{0, \dots, H\}$ and we have for $j = 1, \dots, N$,

$$\int_{I_j^N} g_c(x) dx = \int_{I_j^N} f_c(x) dx = \frac{1}{N},$$

which yields

$$\begin{aligned} M_N &= p_0 \left(\sum_{j=2}^{N-1} \max_{x \in I_j^N}(x^k) \int_{I_j^N} g_c(x) dx + \int_{I_1^N \cup I_N^N} \max \left(\frac{Cx^k}{|x|^{k+1+\xi}}, 0 \right) dx \right) + \sum_{h=1}^H p_h \theta_h^k \\ &\geq p_0 \left(\sum_{j=1}^N \int_{I_j^N} x^k g_c(x) dx \right) + \sum_{h=1}^H p_h \theta_h^k = \int_{\mathbb{R}} x^k g(x) dx = \mathbb{E}_g[X^k]. \end{aligned}$$

Similarly, we have

$$\begin{aligned} m_N &= p_0 \left(\sum_{j=2}^{N-1} \min_{x \in I_j^N}(x^k) \int_{I_j^N} g_c(x) dx \right) + \sum_{h=1}^H p_h \theta_h^k \\ &\leq \int_{\mathbb{R}} x^k g(x) dx = \mathbb{E}_g[X^k], \end{aligned}$$

and the proof is complete. $\square$

We require the use of the Dominated Convergence Theorem [106] to show that the upper bound M_{2^u} converges to $\mathbb{E}_g[X^k]$. In the next lemma, we construct an integrable dominating function $v(x)$.

Lemma 3.C.1. Suppose Assumption 3.2 holds. For $i \geq 1$, let the intervals $J_i = (a_{i-1}, a_i]$, $J_{-i} = (a_{-i}, a_{-i+1}]$ and $J_0 = \emptyset$ be defined such that

$$\int_{-\infty}^{a_0} f_c(x) dx = \frac{1}{2} \quad \text{and} \quad \int_{J_i} f_c(x) dx = \int_{J_{-i}} f_c(x) dx = \frac{1}{2^{i+1}}.$$

Let $v(x) = \sum_{i=-\infty}^{\infty} \max_{x \in J_i} \{|x|^k\} \mathbf{1}_{J_i}(x) f_c(x)$, then $v(x)$ is integrable $\int_{\mathbb{R}} v(x) dx < \infty$

Proof. It is either the case that $a_0 < a_i \leq 0$ for all $i \in \mathbb{N}$ or there exists an n such that $a_i > 0$ for all $i > n$. For the prior case, since $a_0 < a_i \leq 0$ we have for all $i \geq 1$, $0 \leq \max_{x \in J_i} \{|x|^k\} \leq |a_0|^k$. Thus

$$\int_{a_0}^{\infty} v(x) dx = \sum_{i=1}^{\infty} \max_{x \in J_i} \{|x|^k\} \frac{1}{2^{i+1}} < \infty.$$

For the latter case, in order show that

$$\int_{a_0}^{\infty} v(x) dx < \infty,$$

we derive an upperbound for $\max_{x \in J_i} \{|x|^k\}$ so that $\sum_{i=n+1}^{\infty} \max_{x \in J_i} \{|x|^k\} \frac{1}{2^{i+1}} < \infty$.

Using the assumption that $f_c(x) \leq C|x|^{-k-1-\xi}$ and $a_i > 0$ for $i > n$, we obtain

$$\begin{aligned} 2^{-i-1} &= \sum_{j=i}^{\infty} \frac{1}{2^{j+1}} = \sum_{j=i}^{\infty} \int_{J_j} f_c(x) dx \\ &= \int_{a_i}^{\infty} f_c(x) dx \\ &\leq \int_{a_i}^{\infty} C|x|^{-k-1-\xi} dx \\ &= \frac{C}{k+\xi} a_i^{-k-\xi}. \end{aligned} \tag{3.31}$$

Using (3.31), we obtain

$$2^{-i-1} \leq \frac{C}{k+\xi} a_i^{-k-\xi}.$$

Noting that $a_i > 0$ for $i \geq n+1$, we have an upper bound for a_i

$$a_i \leq \left(\frac{C}{k+\xi} \right)^{\frac{1}{k+\xi}} (2^{i+1})^{\frac{1}{k+\xi}} \quad \text{for } i \geq n+1.$$

Thus, we have an upper bound for $\max_{x \in J_i} \{|x|^k\}$ for $i \geq n+1$,

$$\max_{x \in J_i} \{|x|^k\} = a_{i+1}^k \leq \left(2^{\frac{k}{k+\xi}} \right)^{i+1} \left(\frac{C}{k+\xi} \right)^{\frac{k}{k+\xi}} \quad \text{for } i \geq n+1. \quad (3.32)$$

Using the bound in (3.32), we are able to bound the integral $\int_{a_0}^{\infty} v(x) dx$ by

$$\begin{aligned} & \int_{a_0}^{a_{n+1}} v(x) dx + \int_{a_{n+1}}^{\infty} v(x) dx \\ &= \int_{a_0}^{a_{n+1}} v(x) dx + \sum_{i=n+1}^{\infty} \max_{x \in J_i} \{|x|^k\} 2^{-i-1} \\ &\leq \int_{a_0}^{a_{n+1}} v(x) dx + \sum_{i=n+1}^{\infty} \left(2^{\frac{k}{k+\xi}} \right)^{i+1} \left(\frac{C}{k+\xi} \right)^{\frac{k}{k+\xi}} 2^{-i-1} \\ &\leq \int_{a_0}^{a_{n+1}} v(x) dx + \sum_{i=n+1}^{\infty} \left(2^{-\frac{\xi}{k+\xi}} \right)^{i+1} \left(\frac{C}{k+\xi} \right)^{\frac{k}{k+\xi}}. \end{aligned} \quad (3.33)$$

Since $v(x)$ is bounded on the closed interval $[a_0, a_{n+1}]$, $\int_{a_0}^{a_{n+1}} v(x) dx$ is finite. Furthermore, there exist a positive real number B_3 such that

$$\left(\frac{C}{k+\xi} \right)^{\frac{k}{k+\xi}} < B_3 \quad \text{for } k \geq n+1.$$

Therefore, we bound the summation in (3.33) by

$$\sum_{i=n+1}^{\infty} \left(2^{-\frac{\xi}{k+\xi}} \right)^{i+1} \left(\frac{C}{k+\xi} \right)^{\frac{k}{k+\xi}} < B_3 \sum_{i=n+1}^{\infty} \left(2^{-\frac{\xi}{k+\xi}} \right)^{i+1}.$$

Thus, we conclude that $\int_{a_0}^{\infty} v(x)dx < \infty$. A similar argument can be used to show that $\int_{-\infty}^{a_0} v(x)dx < \infty$. Therefore, we have $\int_{\mathbb{R}} v(x)dx < \infty$. $\square$

Lemma 3.C.2. Suppose Assumption 2 holds. For any $j \in \{2, \dots, 2^u - 1\}$, there exists an integer i such that $I_j^{2^u} \subseteq J_i$. Furthermore, for any positive integer u , we have

$$\left| \sum_{j=2}^{2^u-1} \max_{x \in I_j^{2^u}} (x^k) f_c(x) \mathbf{1}_{I_j}(x) \right| \leq v(x) \quad \text{for } x \in \mathbb{R}. \quad (3.34)$$

Proof. We will show that for any $j \in \{2, \dots, 2^u - 1\}$, there exists an integer i such that $I_j^{2^u} \subseteq J_i$ by induction on u . For $u = 2$, we can check that $I_2^{2^2} \subseteq J_{-1}, I_3^{2^2} \subseteq J_1$. Suppose the statement is true for $u = u_0$. For $u = u_0 + 1$, the intervals $\{I_j^{2^{u_0+1}} | j = 1, \dots, 2^{u_0+1}\}$ is a refinement of $\{I_j^{2^{u_0}} | j = 1, \dots, 2^{u_0}\}$. Hence, for each $j \in \{3, \dots, 2^{u_0+1} - 2\}$, there exist an integer i such that $I_j^{2^{u_0+1}} \subseteq J_i$. It remains for us to check that $I_2^{2^{u_0+1}} \subseteq J_{-u_0}$ and that $I_{2^{u_0+1}-1}^{2^{u_0+1}} \subseteq J_{u_0}$. This is the case because the end-points of the intervals a_i corresponds to point $\int_{a_i}^{\infty} f_c(x)dx = \frac{1}{2^{i+1}}$ if $i > 0$ and $\int_{-\infty}^{a_i} f_c(x)dx = \frac{1}{2^{-i+1}}$ if $i < 0$. So $\{J_j\}$ partitions the real line into smaller and smaller intervals towards infinity. For a fixed u , $I_2^{2^u} \subseteq J_{-(u-1)}$ and $I_{2^u-1}^{2^u} \subseteq J_{u-1}$. By mathematical induction, for any $j \in \{2, \dots, 2^u - 1\}$ there exist i such that $I_j^{2^u} \subseteq J_i$ for $u \geq 2$. For a fixed i , for any j such that $I_j^{2^u} \subseteq J_i$, we have

$$\max_{x \in I_j^{2^u}} (|x|^k) \mathbf{1}_{I_j^{2^u}}(x) \leq \max_{x \in J_i} (|x|^k) \mathbf{1}_{J_i}(x) \quad \text{for any } x \in \mathbb{R}.$$

For each of the summands in (3.34), we have the following bound

$$\begin{aligned} \left| \max_{x \in I_j^{2^u}} (x^k) \mathbf{1}_{I_j^{2^u}}(x) f_c(x) \right| &\leq \max_{x \in I_j^{2^u}} (|x|^k) \mathbf{1}_{I_j^{2^u}}(x) f_c(x) \\ &\leq \max_{x \in J_i} (|x|^k) \mathbf{1}_{J_i}(x) f_c(x) \quad \text{for any } x \in \mathbb{R}. \end{aligned}$$

Putting everything together, we obtain

$$\left| \sum_{j=2}^{2^u-1} \max_{x \in I_j^{2^u}} (x^k) f_c(x) \mathbf{1}_{I_j^{2^u}}(x) \right| = \sum_{j=2}^{2^u-1} \left| \max_{x \in I_j^{2^u}} (x^k) f_c(x) \mathbf{1}_{I_j^{2^u}}(x) \right|$$

$$\leq v(x) \text{ for } x \in \mathbb{R}$$

□

Proposition 3.C.4. Under the setup of Theorem 3.C.1, we have $\lim_{u \rightarrow \infty} M_{2^u} = \lim_{u \rightarrow \infty} m_{2^u} = \mathbb{E}_f [X^{k+1}]$.

Proof. We define the upper and lower bound functions $u_N(x)$ and $l_N(x)$ as

$$\begin{aligned} u_N(x) = & \max \left(\frac{Cx^k}{|x|^{-k-1-\xi}}, 0 \right) \mathbf{1}_{I_1^N}(x) \\ & + \sum_{i=2}^N \max_{x \in I_i^N} (x^k) f_c(x) \mathbf{1}_{I_i^N}(x) \\ & + \max \left(\frac{Cx^k}{|x|^{-k-1-\xi}}, 0 \right) \mathbf{1}_{I_N^N}(x), \end{aligned} \quad (3.35)$$

$$l_N(x) = \sum_{i=2}^N \min_{x \in I_i^N} (x^k) f_c(x) \mathbf{1}_{I_i^N}(x), \quad (3.36)$$

Note that $p_0 \int_{\mathbb{R}} l_N(x) dx + \sum_{h=1}^H p_h \theta_h^k = m_N$ and $p_0 \int_{\mathbb{R}} u_N(x) dx + \sum_{h=1}^H p_h \theta_h^k = M_N$. Furthermore, l_N and u_N converges pointwise to $x^k f_c(x)$ as N tends to infinity. Since $x^k f_c(x)$ is integrable and by the Monotone Convergence Theorem [106], we have

$$\lim_{u \rightarrow \infty} m_{2^u} = \lim_{u \rightarrow \infty} p_0 \int_{\mathbb{R}} l_{2^u}(x) dx + \sum_{h=1}^H p_h \theta_h^k = \int_{\mathbb{R}} x^k f_c(x) dx = \mathbb{E}_f [X^k]. \quad (3.37)$$

From Lemma 3.C.1 and 3.C.2, $\sum_{j=2}^{2^u-1} \max_{x \in I_j^{2^u}} (x^{k+1}) f_c(x)$ is dominated by an integrable function $v(x)$. By applying the Lebesgue Dominated Convergence Theorem [106], we obtain

$$\lim_{u \rightarrow \infty} \int_{\mathbb{R}} \sum_{j=2}^{2^u-1} \max_{x \in I_j^{2^u}} (x^k) f_c(x) dx = \int_{\mathbb{R}} x^k f_c(x) dx.$$

Since

$$\lim_{u \rightarrow \infty} \int_{I_1^{2^u} \cup I_{2^u}^{2^u}} \max \left(\frac{Cx^k}{|x|^{k+1+\xi}}, 0 \right) dx = 0,$$

we have

$$\begin{aligned}
\lim_{u \rightarrow \infty} M_{2^u} &= \lim_{u \rightarrow \infty} p_0 \left(\int_{\mathbb{R}} \sum_{j=2}^{2^u-1} \max(x^k) f(x) dx \right. \\
&\quad \left. + \int_{I_1^{2^u} \cup I_{2^u}^{2^u}} \max \left(\frac{C x^k}{|x|^{k+1+\xi}}, 0 \right) dx \right) + \sum_{h=1}^H p_h \theta_h^k \\
&= \int_{\mathbb{R}} x^k f(x) dx + 0 = \mathbb{E}_f [X^k].
\end{aligned} \tag{3.38}$$

□

3.C.3 Simulations

We consider the case where $g = \mathcal{N}(0, 0.5)$, which differs in the second moment from $f = \mathcal{N}(0, 1)$. and we compare the ADD-ARL performance of using $N = N_{\text{exact}}$ by applying Proposition 3.C.1 against using $N = N_{\text{approx}}$ obtained by Algorithm 1. Since the number of intersections of the cdfs of $\mathbb{P}^f$ and $\mathbb{P}^g$ is 3, $\mathbb{P}^g \in D(\mathbb{P}^f, 3)$. Furthermore, since both $\mathbb{P}^f$ and $\mathbb{P}^g$ are absolutely continuous symmetric distributions, $\mathbb{P}^g \notin D(\mathbb{P}^f, 2)$. Therefore the smallest parameter N such that $\mathbb{P}^g \in D(\mathbb{P}^f, N)$ is $N_{\text{exact}} = 3$. On the other hand, applying Algorithm 1 gives us an estimate for the parameter to be $N_{\text{approx}} = 25$, if we assume $k = 2$, $\epsilon = 0.5$, $C = 1.9$ and $\xi = 4$. Figure 3.9a shows the upper bound M_N and lower bound m_N in Algorithm 1 as the parameter N varies. Figure 3.9 shows the ADD versus ARL performance for these values of N using 5000 Monte Carlo trials.

FIG. 3.9: (a) Graph of the upper and lower bounds computed in Algorithm 1. (b) Comparison of ADD vs ARL performance using $N = N_{\text{exact}}$ and $N = N_{\text{approx}}$.

Chapter 4

QCD in the presence of a nuisance change

In this chapter, we discuss the QCD problem when a nuisance change may occur. In Section 4.1, we present our signal model and problem formulation. We propose the W-SGLR stopping time and derive the theoretical properties of our test statistics in Section 4.2. In Section 4.3, we discuss a modification of the proposed stopping time when the post-change distribution belongs to a parametrized family. We present numerical simulations and experiments on a real dataset to illustrate the performance of our proposed stopping time in Section 4.4.

Notations: For the rest of this chapter, we let $\mathbb{P}_{\nu_n, \nu_c}$ and $\mathbb{E}_{\nu_n, \nu_c}$ be the probability measure and mathematical expectation when the nuisance change point is at ν_n , and the critical change point is at ν_c , respectively. Here, the subscript n denotes the word nuisance in this chapter.

4.1 Problem formulation

In many applications, the statistical distribution of the observed signal may undergo different changes over time. For example, in the application of fault detection

in motor bearings [35], we aim to raise an alarm as soon as possible after a bearing fault has occurred (critical change). This is done by monitoring the accelerometer readings from the motor to detect any changes in the signal statistics. However, the accelerometer readings are also affected by non-critical or nuisance changes like variation in the motor-load of the bearing. It would be undesirable if we declare that a fault has taken place whenever the motor-load changes. This motivates a need to define a change-point model that allows both critical and nuisances changes and to develop change detection techniques that can effectively ignore nuisance changes while efficiently identifying critical changes.

In this chapter, we assume that the signals observed, $X_1, X_2, \dots$, may undergo two types of change: a critical change at $\nu_c \geq 0$ and a nuisance change at $\nu_n \geq 0$. Both the critical and nuisance change points are unknown *a priori*. We are interested in detecting the critical change while the nuisance change is not of interest. Let $\mathbb{P}^f, \mathbb{P}^{f_n}, \mathbb{P}^g, \mathbb{P}^{g_n}$ be probability distributions. At each time t , we let $\mathbb{P}^{h_{\nu_n, \nu_c, t}}$ to be the distribution that generates the observation X_t when the nuisance change point is at ν_n and the critical change point is at ν_c :

$$h_{\nu_n, \nu_c, t} = \begin{cases} f & \text{if } t < \min\{\nu_c, \nu_n\}, \\ f_n & \text{if } \nu_n \leq t < \nu_c, \\ g & \text{if } \nu_c \leq t < \nu_n, \\ g_n & \text{if } \max\{\nu_c, \nu_n\} \leq t. \end{cases} \quad (4.1)$$

Thus, in our model (cf. Fig. 4.1), $\mathbb{P}^f$ is the pre-change distribution. If $\nu_c < \nu_n < \infty$, the signal distribution first changes to $\mathbb{P}^g$ at ν_c and then to $\mathbb{P}^{g_n}$ at ν_n . If $\nu_n < \nu_c < \infty$, the distribution first changes to $\mathbb{P}^{f_n}$ at ν_n and then to $\mathbb{P}^{g_n}$ at ν_c . If $\nu_n = \nu_c$, then the distribution changes from $\mathbb{P}^f$ to $\mathbb{P}^{g_n}$ at the common change point.

The sequence of observations $X_1, X_2, \dots$ is a sequence of random variables satisfying $X_t \sim h_{\nu_n, \nu_c, t}$ where $\{X_t\}_{t \in \mathbb{N}}$ are mutually independent given ν_n, ν_c . The quickest change detection problem is to detect the critical change ν_c through observing $X_1, X_2, \dots$, as quickly as possible while ignoring the nuisance change and keeping

FIG. 4.1: Evolution of the signal distribution in our model. In this example, changes in mean are considered nuisance and changes in variance are considered critical.

the false alarm rate low. In our signal model, the nuisance change also changes the distribution that generates the observations after the critical change point. This creates a dependence between the nuisance change point and the distribution after the critical change point. Our formulation is different from assuming composite pre-change and post-change distribution families [113] since the nuisance change leads to non-stationarity in the distribution of X_t before or after the critical change, depending on whether the nuisance change occurs before or after the critical change, respectively.

In a typical sequential change detection procedure, at each time t , a test statistic $S(t)$ is computed based on the currently available observations $X_1, \dots, X_t$, and the observer decides that a change has occurred at a stopping time $\tau(b) = \inf\{t : S(t) \geq b\}$.

In the traditional QCD framework [7], the rate of false alarms is quantified by the mean time between false alarms. Since the nuisance change-point affects the distributions generating the signal, this quantity varies with the nuisance change point. In this chapter, we consider the worst-possible rate as the nuisance change point varies by considering the smallest mean time between false alarms for all possible nuisance change points. A similar generalization can be made to quantify

the detection delay by taking the largest detection delay over all possible nuisance change points.

Mathematically, our QCD problem can be formulated as a minimax problem similar to Lorden's formulation [7], where we seek a stopping time that minimizes the WADD subject to an ARL constraint:

$$\begin{aligned} \min_{\tau} \quad & \text{WADD}(\tau) \\ \text{s. t.} \quad & \text{ARL}(\tau) \geq \gamma, \end{aligned} \tag{4.2}$$

where γ is a predefined threshold, τ is a stopping time with respect to (w.r.t.) the filtration $\{\sigma(X_1, X_2, \dots, X_t) : t \geq 0\}$,

$$\text{WADD}(\tau) = \sup_{\nu_n \in \mathbb{N} \cup \{\infty\}} \text{WADD}_{\nu_n}(\tau), \tag{4.3}$$

$$\text{WADD}_{\nu_n}(\tau) = \sup_{\nu_c \geq 1} \text{ess sup } \mathbb{E}_{\nu_n, \nu_c} [(\tau - \nu_c + 1)^+ \mid X_1, \dots, X_{\nu_c-1}], \tag{4.4}$$

$$\text{ARL}(\tau) = \inf_{\nu_n \in \mathbb{N} \cup \{\infty\}} \mathbb{E}_{\nu_n, \infty} [\tau], \tag{4.5}$$

and ess sup is the essential supremum operator. In the next section, we propose a stopping time for (4.2).

A closely related topic is transient change detection (TCD) [114–117] where the change only occurs for a finite period of time and the objective is to detect if such a change has occurred within a predefined window or not instead of detecting the change as quickly as possible. There are two widely adopted methods for the TCD problem, the window-limited CuSum stopping time [56] and the finite moving average (FMA) stopping time [117]. The FMA stopping time has been shown to perform well for the TCD problem, and we will use the FMA stopping time as a comparison in Section 4.4. When $\nu_c < \nu_n$, our system model can be seen to be a generalization of the TCD problem variant where one seeks to detect the transient change as quickly as possible by letting $g_n = f$. In (4.16) below,

we propose a test statistic and stopping time for (4.2). By setting $g_n = f$ and $f_n = f$ in our test statistic, our proposed stopping time reduces to the window-limited CuSum stopping time with pre-change distribution $\mathbb{P}^f$ and post-change distribution $\mathbb{P}^g$.

4.2 Test Statistic for QCD with Nuisance Change

In this section, we derive a test-statistic and stopping time for QCD under a nuisance change. Suppose that we observe the sequence $X_1, X_2, \dots$ and know *a priori* that the nuisance change does not take place (i.e., $\nu_n = \infty$), then Page's CuSum test statistic [6] given as $S_{\text{CuSum}}(t) = \max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{g(X_i)}{f(X_i)}$, can be used and we declare that a critical change has taken place at

$$\tau_{\text{CuSum}}(b) = \inf \{t : S_{\text{CuSum}}(t) \geq b\} \quad (4.6)$$

$$= \inf \left\{ t : \max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{g(X_i)}{f(X_i)} \geq b \right\}, \quad (4.7)$$

where b is a pre-determined threshold. The CuSum test statistics has a convenient recursion $S_{\text{CuSum}}(t+1) = \max \left\{ S_{\text{CuSum}}(t) + \log \frac{g(X_{t+1})}{f(X_{t+1})}, 0 \right\}$, which allows the CuSum stopping time to be implemented efficiently.

If the nuisance change takes places at a time $\nu_n < \infty$ and ν_n is known, a modification of Page's test statistic gives the following:

$$S_{\text{CuSum}}(t) = \max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{h_{\nu_n,1,i}(X_i)}{h_{\nu_n,\infty,i}(X_i)}, \quad (4.8)$$

$$\tau_{\text{CuSum}}(b) = \inf \left\{ t : \max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{h_{\nu_n,1,i}(X_i)}{h_{\nu_n,\infty,i}(X_i)} \geq b \right\}, \quad (4.9)$$

where $h_{\nu_n,1,i}(x)$ and $h_{\nu_n,\infty,i}(x)$ are as defined in (4.1) and are the probability distributions corresponding to the cases where the critical change has already occurred or will never occur, respectively. Similar to the case where $\nu_n = \infty$, the CuSum

test statistics admits a convenient recursion for efficient implementation. Furthermore, for both the cases mentioned above, τ_{CuSum} was shown to be asymptotically optimal by [7].

A naive approach is to utilize four variants of τ_{CuSum} , one for detecting for a change in each of the cases: from f to f_n , from f to g , from f to g_n , and from f_n to g_n . In the first stage, we monitor for changes from f to either f_n , g or g_n . If a change to f_n is detected, then we monitor for a change from f_n to g_n . The difficulty in such an approach is that any false alarm or miss detection in the first stage propagates to the second stage. We demonstrate that such an approach is suboptimal in Section 4.4.1.

In our problem formulation, the nuisance change-point ν_n is unknown. Replacing ν_n with its maximum likelihood estimator in both the numerator and denominator, we obtain the following GLR test statistic and stopping time:

$$\begin{aligned}\Lambda_{\text{GLR}}(k, t) &= \frac{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j,1,i}(X_i)}{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j,\infty,i}(X_i)} \\ &= \frac{\max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} g(X_i) \prod_{i=j}^t g_n(X_i)}{\max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} f(X_i) \prod_{i=j}^t f_n(X_i)},\end{aligned}\tag{4.10}$$

$$S_{\text{GLR}}(t) = \max_{1 \leq k \leq t+1} \log \Lambda_{\text{GLR}}(k, t),\tag{4.11}$$

$$\tau_{\text{GLR}}(b) = \inf \{t : S_{\text{GLR}}(t) \geq b\}.\tag{4.12}$$

From our simulations in Section 4.4.1, it turns out that (4.12) does not achieve the best trade-off between average detection delay (ADD) and ARL to false alarm over a wide range of threshold values b . Furthermore, its ARL is challenging to characterize theoretically since the GLR test statistic $\Lambda_{\text{GLR}}(k, t)$ is not a likelihood ratio and standard techniques in the QCD literature (e.g., Theorem 6.16 of [18]) cannot be used to analyze its ARL. This is a critical problem for practical applications that require us to pre-determine a suitable threshold b to achieve a desired ARL.

To develop a stopping time with ARL that can be characterized theoretically, we simplify the maximum likelihood estimation in the numerator of (4.10) to be the maximum of only two cases $j = k$ and $j = t + 1$. This gives us the Simplified GLR (SGLR) test statistic and stopping time as follows:

$$\Lambda_{\text{SGLR}}(k, t) = \frac{\max \left\{ \prod_{i=k}^t g(X_i), \prod_{i=k}^t g_n(X_i) \right\}}{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j, \infty, i}(X_i)} \quad (4.13)$$

$$S_{\text{SGLR}}(t) = \max_{1 \leq k \leq t+1} \log \Lambda_{\text{SGLR}}(k, t), \quad (4.14)$$

$$\tau_{\text{SGLR}}(b) = \inf \{t : S_{\text{SGLR}}(t) \geq b\}. \quad (4.15)$$

Unlike the CuSum test statistic, the SGLR test statistic does not have a convenient recursion. Any implementation of the SGLR stopping time would require computational resources that increases with the number of samples observed. The requirement on computational resources would be a significant limitation for many practical applications. To limit the computational resources required, in the same spirit as [9], we propose the Window-Limited SGLR (W-SGLR) test statistic and stopping time as follows:

$$S_{\text{W-SGLR}}(t) = \max_{t-m_b \leq k \leq t+1} \log \Lambda_{\text{SGLR}}(k, t), \quad (4.16)$$

$$\tau_{\text{W-SGLR}}(b) = \inf \{t : S_{\text{W-SGLR}}(t) \geq b\}, \quad (4.17)$$

where the window size m_b is chosen such that

$$\liminf_{b \rightarrow \infty} \frac{m_b}{b} > I^{-1} \text{ and } \log m_b = o(b), \quad (4.18)$$

with

$$I = \min \{D(g \| f), D(g \| f_n), D(g_n \| f), D(g_n \| f_n)\}, \quad (4.19)$$

and $o(b)$ denoting a term that goes to zero as $b \rightarrow \infty$. Window-limited test statistics were first introduced by [118]. The paper [9] further discussed their properties and

the choice of window size and thresholds. We make the following assumption.

Assumption 4.1. The first four moments of $\log \frac{f_n(X)}{f(X)}$ w.r.t. both g and g_n are finite, and $\rho_g, \rho_{g_n} \neq 0$, where we define $\rho_g = \mathbb{E}^g \left[\log \frac{f_n(X)}{f(X)} \right]$, $\sigma_g^2 = \mathbb{E}^g \left[\left(\log \frac{f_n(X)}{f(X)} - \rho_g \right)^2 \right]$, $\omega_g^4 = \mathbb{E}^g \left[\left(\log \frac{f_n(X)}{f(X)} - \rho_g \right)^4 \right]$, $\rho_{g_n} = \mathbb{E}^{g_n} \left[\log \frac{f_n(X)}{f(X)} \right]$, $\sigma_{g_n}^2 = \mathbb{E}^{g_n} \left[\left(\log \frac{f_n(X)}{f(X)} - \rho_{g_n} \right)^2 \right]$ and $\omega_{g_n}^4 = \mathbb{E}^{g_n} \left[\left(\log \frac{f_n(X)}{f(X)} - \rho_{g_n} \right)^4 \right]$.

We require Assumption 4.1 in order to control the tail probabilities of $\log \frac{f_n(X)}{f(X)}$ when deriving the asymptotic bounds of WADD of the W-SGLR stopping time. One general class of distributions that satisfies Assumption 4.1 is the class of natural exponential families.

In Theorem 4.3 of Section 4.2.2, we show that the proposed $\tau_{\text{W-SGLR}}(b)$ is asymptotically optimal as $b \rightarrow \infty$ under Assumption 4.1 and an additional technical assumption. To do that, we first analyze the asymptotic properties of $\tau_{\text{W-SGLR}}$. We let

$$\Lambda(k, t) = \frac{\prod_{i=k}^t g(X_i)}{\max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} f(X_i) \prod_{i=j}^t f_n(X_i)}, \quad (4.20)$$

$$\Lambda_n(k, t) = \frac{\prod_{i=k}^t g_n(X_i)}{\max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} f(X_i) \prod_{i=j}^t f_n(X_i)}, \quad (4.21)$$

and study their properties in Section 4.2.1. Then, using the relationships

$$\Lambda_{\text{SGLR}}(k, t) = \max \{ \Lambda(k, t), \Lambda_n(k, t) \}, \quad (4.22)$$

$$\tau_{\text{SGLR}}(b) = \min \{ \tau(b), \tau_n(b) \}, \quad (4.23)$$

$$\tau_{\text{W-SGLR}}(b) = \min \{ \tilde{\tau}(b), \tilde{\tau}_n(b) \}, \quad (4.24)$$

where

$$\begin{aligned}
\tau(b) &= \inf\{t : \max_{k \leq t} \log \Lambda(k, t) \geq b\}, \\
\tau_n(b) &= \inf\{t : \max_{k \leq t} \log \Lambda_n(k, t) \geq b\}, \\
\tilde{\tau}(b) &= \inf\{t : \max_{t-m_b \leq k \leq t} \log \Lambda(k, t) \geq b\}, \\
\tilde{\tau}_n(b) &= \inf\{t : \max_{t-m_b \leq k \leq t} \log \Lambda_n(k, t) \geq b\},
\end{aligned} \tag{4.25}$$

we finally show the asymptotic optimality of $\tau_{\text{W-SGLR}}$ under mild technical conditions in Section 4.2.2.

4.2.1 Log Likelihood Ratio Growth Rates

In this subsection, we derive properties of Λ and Λ_n as defined in (4.20) and (4.21), respectively. The stopping times $\tau_{\text{SGLR}}(b)$ and $\tau_{\text{W-SGLR}}(b)$ are defined by the first time the test statistics S_{SGLR} and $S_{\text{W-SGLR}}$ cross the threshold b respectively. The rates of growth, $\frac{1}{t-k+1} \log \Lambda(k, t)$ and $\frac{1}{t-k+1} \log \Lambda_n(k, t)$, allow us to understand the detection delay of these stopping times. We show that these rates of growth converge in probability as $t \rightarrow \infty$. In particular, the limit that the rate of growth converges to depends on the sign of ρ_g and ρ_{g_n} .

As the nuisance change point is unknown, the denominator of both Λ and Λ_n contains a maximization of the likelihood $\max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} f(X_i) \prod_{i=j}^t f_n(X_i)$. If the first moment $\rho_g < 0$, the distribution $\mathbb{P}^g$ is closer to the distribution $\mathbb{P}^f$ as compared to $\mathbb{P}^{f_n}$ in the Kullback-Leibler (KL) divergence sense. When the critical change point is at $\nu_c = 1$ and no nuisance change has taken place, we expect the denominator to approach $\prod_{i=k}^t f(X_i)$. Thus, our statistic $\Lambda(k, t)$ can be approximated by $\prod_{i=k}^t \frac{g(X_i)}{f(X_i)}$. A similar argument can be made for Λ_n when $\nu_n = \nu_c = 1$. This observation is made precise in the following two propositions.

Proposition 4.1. Suppose that Assumption 4.1 holds, and $\rho_g < 0$. For any $\nu_c \leq k < \infty$ and $\epsilon > 0$, we have

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\infty, \nu_c} \left(\left| \frac{\log \Lambda(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g(X_i)}{f(X_i)} \right| \geq \epsilon \right) = 0, \quad (4.26)$$

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\infty, \nu_c} \left(\left| \frac{\log \Lambda_n(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g_n(X_i)}{f(X_i)} \right| \geq \epsilon \right) = 0. \quad (4.27)$$

Proof. See Appendix 4.A. □

Proposition 4.2. Suppose that Assumption 4.1 holds, and $\rho_{g_n} > 0$. For any $\nu_c \leq k < \infty$, $\nu_n < \infty$, and $\epsilon > 0$, we have

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{\log \Lambda(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g(X_i)}{f_n(X_i)} \right| \geq \epsilon \right) = 0, \quad (4.28)$$

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{\log \Lambda_n(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g_n(X_i)}{f_n(X_i)} \right| \geq \epsilon \right) = 0. \quad (4.29)$$

Proof. See Appendix 4.A. □

Using Propositions 4.1 and 4.2 together with the weak law of large numbers, we obtain the following result.

Theorem 4.1. Suppose that Assumption 4.1 holds, $\rho_g = D(g \| f) - D(g \| f_n) < 0$, and $\rho_{g_n} = D(g_n \| f) - D(g_n \| f_n) > 0$. For any $\nu_c \leq k < \infty$,

$$\begin{aligned} \frac{\log \Lambda(k, t)}{t - k + 1} &\xrightarrow{\mathbb{P}_{\infty, \nu_c}} D(g \| f), \\ \frac{\log \Lambda_n(k, t)}{t - k + 1} &\xrightarrow{\mathbb{P}_{\infty, \nu_c}} D(g \| f) - D(g \| g_n), \end{aligned}$$

as $t \rightarrow \infty$. Furthermore, for any $\nu_n < \infty$,

$$\begin{aligned} \frac{\log \Lambda(k, t)}{t - k + 1} &\xrightarrow{\mathbb{P}_{\nu_n, \nu_c}} D(g_n \| f_n) - D(g_n \| g), \\ \frac{\log \Lambda_n(k, t)}{t - k + 1} &\xrightarrow{\mathbb{P}_{\nu_n, \nu_c}} D(g_n \| f_n), \end{aligned}$$

as $t \rightarrow \infty$.

In Theorem 4.1, we have assumed that $\rho_g < 0$ and $\rho_{g_n} > 0$. If we vary the signs of ρ_g and ρ_{g_n} , a similar argument to that provided in Theorem 4.1 gives us the following result.

Theorem 4.2. Suppose that Assumption 4.1 holds. For any $\nu_c \leq k < \infty$ and $\nu_n < \infty$, we have the following convergences in probability as $t \rightarrow \infty$ shown in the table below.

Sign of moment		under $\mathbb{P}_{\infty, \nu_c}$	under $\mathbb{P}_{\nu_n, \nu_c}$
ρ_g	ρ_{g_n}	$\frac{\log \Lambda(k, t)}{t-k+1}$	$\frac{\log \Lambda_n(k, t)}{t-k+1}$
>0	>0	$D(g \parallel f_n)$	$D(g_n \parallel f_n)$
>0	<0	$D(g \parallel f_n)$	$D(g_n \parallel f)$
<0	>0	$D(g \parallel f)$	$D(g_n \parallel f_n)$
<0	<0	$D(g \parallel f)$	$D(g_n \parallel f)$

Theorem 4.2 gives us the average rate of growth of the statistics $\log \Lambda(k, t)$ and $\log \Lambda_n(k, t)$. Since I in (4.19) is the minimum of the growth rates in Theorem 4.2, we see that the average growth rate of these statistics is at least I regardless of the signs of ρ_g and ρ_{g_n} . This suggests that the WADD of $\tau_{\text{W-SGLR}}(b)$ grows linearly with respect to b with a gradient bounded above by I . This observation is made precise in the next subsection.

4.2.2 Conditions for Asymptotic Optimality

In this subsection, we establish the asymptotic WADD-ARL trade-off under Assumption 4.1 and provide a sufficient condition for $\tau_{\text{W-SGLR}}$ to be asymptotically optimal. In particular, we show that $\tau_{\text{W-SGLR}}$ is asymptotically optimal if in addition to Assumption 4.1, the following assumption holds.

Assumption 4.2. The KL divergences $D(g \parallel f)$, $D(g \parallel f_n)$, $D(g_n \parallel f)$ and $D(g_n \parallel f_n)$ satisfy

$$D(g \parallel f_n) > \min \{D(g \parallel f), D(g_n \parallel f), D(g_n \parallel f_n)\}.$$

Assumption 4.2 essentially says that g cannot be too similar to f_n , which makes intuitive sense as otherwise it is difficult to distinguish the critical change from the nuisance change (see Fig. 4.1). A sufficient condition for Assumption 4.2 is $\rho_g < 0$ as assumed in Theorem 4.1. For example, in the problem of spectrum sensing in cognitive radio [119], we are often interested in detecting a variance change of a signal generated by independent Gaussian distributions. Furthermore, in many signal processing applications, a change in mean may be due to sensor drift as a result of long duration monitoring. This change in mean is usually not of interest and interferes with the actual signal processing task [120–122]. A typical signal model of this type is given by $f = \mathcal{N}(\mu_0, \sigma_0^2)$, $f_n = \mathcal{N}(\mu_1, \sigma_0^2)$, $g = \mathcal{N}(\mu_0, \sigma_1^2)$, and $g_n = \mathcal{N}(\mu_1, \sigma_1^2)$ with $\mu_0, \mu_1 \in \mathbb{R}$, $\mu_0 \neq \mu_1$, $\sigma_0, \sigma_1 \in \mathbb{R}_{>0}$. While Assumption 4.2 may seem artificial, it is shown in Lemma 4.3 that this model satisfies $\rho_g < 0$ and hence Assumption 4.2. In this case, the W-SGLR stopping time achieves asymptotic optimality.

We use techniques from the proof of Theorem 6.16 in [18] to obtain a lower bound for the ARL of τ_{SGLR} in (4.15). Since $\tau_{\text{W-SGLR}} \geq \tau_{\text{SGLR}}$, the same lower bound also applies for the ARL of $\tau_{\text{W-SGLR}}$ in (4.17). In the previous subsection, we have shown that the rate of growth of the statistics Λ and Λ_n converge to constants as $t \rightarrow \infty$. This means that, asymptotically, Λ and Λ_n grow linearly w.r.t. t . Heuristically, this implies that the WADD of the stopping times, $\tau(b)$ and $\tau_n(b)$, grow linearly w.r.t. the threshold b in (4.25).

Lemma 4.1 derives an upper-bound for the probability of a false alarm for a stopping time related to $\tau_{\text{W-SGLR}}(b)$. Following Theorem 6.16 in [18], this upper bound then yields a lower bound for the ARL of $\tau_{\text{W-SGLR}}(b)$ in Theorem 4.3.

Lemma 4.1. For $b > 0$, let η^k, η_n^k be stopping times defined by

$$\begin{aligned}\eta^k &= \inf\{t \geq k : \log \Lambda(k, t) \geq b\}, \\ \eta_n^k &= \inf\{t \geq k : \log \Lambda_n(k, t) \geq b\},\end{aligned}$$

so that $\tau(b) = \inf_{k \geq 1} \eta^k$ and $\tau_n(b) = \inf_{k \geq 1} \eta_n^k$ in (4.25). For any $\nu_n \in \mathbb{N} \cup \{\infty\}$, we have

$$\mathbb{P}_{\nu_n, \infty}(\eta^1 < \infty) \leq e^{-b} \quad \text{and} \quad \mathbb{P}_{\nu_n, \infty}(\eta_n^1 < \infty) \leq e^{-b}, \quad (4.30)$$

and

$$\mathbb{E}_{\nu_n, \infty}[\tau_{\text{W-SGLR}}(b)] \geq \mathbb{E}_{\nu_n, \infty}[\tau_{\text{SGLR}}(b)] \geq \frac{1}{2}e^b. \quad (4.31)$$

Proof. See Appendix 4.B. □

The next lemma checks that our proposed stopping time satisfies the assumption required in [9] to relate the asymptotic upper-bound for the WADD to the threshold b in Proposition 4.3.

Lemma 4.2. Suppose that Assumption 4.1 holds. For any $\delta > 0$, we have

- (i) $\lim_{t \rightarrow \infty} \sup_{\nu_n \in \mathbb{N}, 1 \leq \nu_c \leq k} \mathbb{P}_{\nu_n, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k+t-1) - I \leq -\delta \right) = 0$, and
- (ii) $\lim_{t \rightarrow \infty} \sup_{1 \leq \nu_c \leq k} \mathbb{P}_{\infty, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k+t-1) - I \leq -\delta \right) = 0$.

Proof. See Appendix 4.C. □

Proposition 4.3. Suppose that Assumption 4.1 holds. There exists a B such that for all $b \geq B$, we have

- (i) $\sup_{\nu_n, \nu_c \geq 1} \text{ess sup } \mathbb{E}_{\nu_n, \nu_c} [(\tilde{\tau}_n(b) - \nu_c + 1)^+ \mid X_1, \dots, X_{\nu_c-1}] \leq (I^{-1} + o(1))b$, and
- (ii) $\sup_{\nu_c \geq 1} \text{ess sup } \mathbb{E}_{\infty, \nu_c} [(\tilde{\tau}(b) - \nu_c + 1)^+ \mid X_1, \dots, X_{\nu_c-1}] \leq (I^{-1} + o(1))b$.

Proof. See Appendix 4.D. □

Finally, we show the asymptotic optimality of $\tau_{\text{W-SGLR}}$ in the following result.

Theorem 4.3. Suppose that Assumption 4.1 holds. For any $b > 0$,

$$\text{ARL}(\tau_{\text{W-SGLR}}(b)) \geq \frac{1}{2}e^b, \quad (4.32)$$

$$\text{WADD}(\tau_{\text{W-SGLR}}(b)) \leq (I^{-1} + o(1))b, \quad (4.33)$$

where $o(1)$ is a term going to zero as $b \rightarrow \infty$. Furthermore, if Assumption 4.2 holds, then the stopping time $\tau_{\text{W-SGLR}}(b)$ is asymptotically optimal for the problem (4.2) as $b \rightarrow \infty$.

Proof. See Appendix 4.E. □

In Theorem 4.3, we have shown that $\tau_{\text{W-SGLR}}$ is asymptotically optimal under Assumption 4.1 and Assumption 4.2. In the next lemma, we derive sufficient conditions for Assumption 4.2 when f, f_n, g, g_n belong to an exponential family.

Lemma 4.3. Suppose that $f, f_n, g, g_n \in \{\phi : \phi(x) = h(x) \exp(\sum_{i=1}^s B_i(\theta)T_i(x) - A(\theta))\}$, an exponential family of distributions on $\mathbb{R}^N$ with parameters $\theta = \theta_f, \theta_{f_n}, \theta_g, \theta_{g_n}$, respectively. Here, $T_i \in \mathbb{R}^N \times \mathbb{R}$ and $A, B_i \in \mathbb{R}^M \times \mathbb{R}$ for $i = 1, \dots, s$. If any of the following inequalities hold:

$$A(\theta_{f_n}) - A(\theta_f) - \sum_{i=1}^s (B_i(\theta_{f_n}) - B_i(\theta_f)) \mathbb{E}^g[T_i(X)] > 0, \quad (4.34)$$

$$A(\theta_{f_n}) - A(\theta_g) - \sum_{i=1}^s (B_i(\theta_{f_n}) - B_i(\theta_g)) \mathbb{E}^g[T_i(X)] > A(\theta_f) - A(\theta_{g_n}) - \sum_{i=1}^s (B_i(\theta_f) - B_i(\theta_{g_n})) \mathbb{E}^{g_n}[T_i(X)], \quad (4.35)$$

$$A(\theta_{f_n}) - A(\theta_g) - \sum_{i=1}^s (B_i(\theta_{f_n}) - B_i(\theta_g)) \mathbb{E}^g[T_i(X)] > A(\theta_{f_n}) - A(\theta_{g_n}) - \sum_{i=1}^s (B_i(\theta_{f_n}) - B_i(\theta_{g_n})) \mathbb{E}^{g_n}[T_i(X)], \quad (4.36)$$

then Assumption 4.2 holds.

In particular, if $f = \mathcal{N}(\mu_0, \sigma_0^2)$, $f_n = \mathcal{N}(\mu_1, \sigma_0^2)$, $g = \mathcal{N}(\mu_0, \sigma_1^2)$, and $g_n = \mathcal{N}(\mu_1, \sigma_1^2)$ with $\mu_0, \mu_1 \in \mathbb{R}$, $\mu_0 \neq \mu_1$, $\sigma_0, \sigma_1 \in \mathbb{R}_{>0}$, and $\sigma_0 \neq \sigma_1$, Assumption 4.2 holds.

Proof. To show that (4.34) implies Assumption 4.2, we rearrange the terms on the left-hand side (LHS) of (4.34) to obtain

$$\mathbb{E}^g \left[\log \frac{f_n(X)}{f(X)} \right] = \mathbb{E}^g \left[\log \frac{h(X) \exp(\sum_{i=1}^s B_i(\theta_{f_n}) T_i(X) - A(\theta_{f_n}))}{h(X) \exp(\sum_{i=1}^s B_i(\theta_f) T_i(X) - A(\theta_f))} \right] < 0. \quad (4.37)$$

This implies that

$$D(g \parallel f) < D(g \parallel f_n), \quad (4.38)$$

and hence Assumption 4.2 holds. A similar argument shows that (4.35) and (4.36) imply $D(g_n \parallel f) < D(g \parallel f_n)$ and $D(g_n \parallel f_n) < D(g \parallel f_n)$, respectively.

If $f = \mathcal{N}(\mu_0, \sigma_0^2)$, $f_n = \mathcal{N}(\mu_1, \sigma_0^2)$, $g = \mathcal{N}(\mu_0, \sigma_1^2)$, and $g_n = \mathcal{N}(\mu_1, \sigma_1^2)$ with $\mu_0, \mu_1 \in \mathbb{R}$, $\mu_0 \neq \mu_1$, $\sigma_0, \sigma_1 \in \mathbb{R}_{>0}$, and $\sigma_0 \neq \sigma_1$, we can define $\theta_f = [\mu_0, \sigma_0^2]$, $\theta_{f_n} = [\mu_1, \sigma_0^2]$, $\theta_g = [\mu_0, \sigma_1^2]$, $\theta_{g_n} = [\mu_1, \sigma_1^2]$, with the functions $B_1(\mu, \sigma^2) = \mu/\sigma^2$, $B_2(\mu, \sigma^2) = \frac{-1}{2\sigma^2}$, $T_1[X] = X$, $T_2[X] = X^2$ and $A(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \log \sigma$. The LHS of (4.34) becomes $\frac{\mu_1^2 - \mu_0^2}{\sigma_0^2} - \left(\frac{\mu_1 - \mu_0}{\sigma_0} \right) \mu_0$. Simplifying, the LHS of (4.34) becomes $\frac{(\mu_1 - \mu_0)^2}{2\sigma_0^2}$. Thus, for any $\mu_1 \neq \mu_0$ and $\sigma_0, \sigma_1 \in \mathbb{R}_{>0}$, the inequality (4.34) holds. The proof is now complete. $\square$

4.3 Parametrized Families of Post-Change Distributions

In many applications, the post-change distribution $\mathbb{P}^g$ and nuisance post-change distribution $\mathbb{P}^{g_n}$ may contain unknown parameters. In this section, we modify τ and τ_n in (4.25) to obtain a Generalized Likelihood Ratio Test (GLRT)-based stopping time $\hat{\tau}_{\text{W-SGLR}}$ for the following signal model: Let $\Theta \subseteq \mathbb{R}^d$ be a set with non-empty interior and $X_1, X_2, \dots$ be a sequence of independent random variables

satisfying: $X_t \sim h_{\nu_n, \nu_c, \theta, \theta_n, t}$ where

$$h_{\nu_n, \nu_c, \theta, \theta_n, t}(\cdot) = \begin{cases} f(\cdot) & \text{if } t < \min\{\nu_c, \nu_n\}, \\ f_n(\cdot) & \text{if } \nu_n \leq t < \nu_c, \\ g(\cdot; \theta) & \text{if } \nu_c \leq t < \nu_n, \\ g_n(\cdot; \theta_n) & \text{if } \max\{\nu_c, \nu_n\} \leq t, \end{cases} \quad (4.39)$$

and $\theta, \theta_n \in \text{Int}(\Theta)$, the interior of Θ . We derive a lower bound for the ARL of $\hat{\tau}_{\text{W-SGLR}}$ under the following assumption.

Assumption 4.3. Θ is a compact d -dimensional sub-manifold of $\mathbb{R}^d$. The pdfs of the post-change distributions $\mathbb{P}^{g(\cdot; \theta)}$ and nuisance post-change distribution $\mathbb{P}^{g_n(\cdot; \theta)}$ are twice continuously differentiable w.r.t. θ .

A commonly used method to handle unknown parameters is to replace the likelihood ratio $\Lambda(k, t)$ with the generalized likelihood ratio. We define the generalized W-SGLR test statistic $\hat{S}_{\text{W-SGLR}}$ as

$$\begin{aligned} \hat{\Lambda}_{\text{W-SGLR}}(k, t, \theta) &= \frac{\max \left\{ \prod_{i=k}^t g(X_i; \theta), \prod_{i=k}^t g_n(X_i; \theta) \right\}}{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j, \infty, \theta, \theta, i}(X_i)}, \\ \hat{S}_{\text{W-SGLR}}(t) &= \max_{t-m_b+1 \leq k \leq t-m'_b} \max_{\theta \in \Theta} \log \hat{\Lambda}(k, t, \theta), \end{aligned} \quad (4.40)$$

where the minimal delay m'_b is required to prevent difficulties of under-determination when performing maximum likelihood estimation of the parameter θ . While (4.40) is commonly used, the maximization over Θ make it difficult to theoretically quantify the ARL of the stopping time $\inf\{t : \hat{S}_{\text{W-SGLR}}(t) \geq b\}$. To work around this problem, we modify the stopping times τ and τ_n as follows. Let $\lambda_{\max}(A)$ denote the largest eigenvalue of the symmetric matrix A . Fix $m'_b \geq 0$. We let

$$\hat{\Lambda}(k, t, \theta) = \frac{\prod_{i=k}^t g(X_i; \theta)}{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j, \infty, \theta, \theta, i}(X_i)}, \quad (4.41)$$

$$\hat{\tau}(b) = \min_{m'_b \leq l \leq m_b} \hat{\eta}_l(b), \quad (4.42)$$

$$\begin{aligned} \hat{\eta}_l(b) = \inf \left\{ t \geq l : \log \hat{\Lambda}(t-l+1, t, \hat{\theta}) \geq b, \right. \\ \left. \sup_{\|\theta - \hat{\theta}\| < 1/\sqrt{b}} \lambda_{\max} \left(-\nabla^2 \log \hat{\Lambda}(t-l+1, t, \theta) \right) \leq b, \right. \\ \left. \text{and } \hat{\theta} = \arg \max_{\theta} \log \hat{\Lambda}(t-l+1, t, \theta) \in \text{Int}(\Theta) \right\}, \end{aligned} \quad (4.43)$$

$$\hat{\Lambda}_n(k, t, \theta) = \frac{\prod_{i=k}^t g_n(X_i; \theta)}{\max_{k \leq j \leq t+1} \prod_{i=k}^t h_{j, \infty, \theta, \theta, i}(X_i)}, \quad (4.44)$$

$$\hat{\tau}_n(b) = \min_{m'_b \leq l \leq m_b} \hat{\eta}_{n,l}(b), \quad (4.45)$$

$$\begin{aligned} \hat{\eta}_{n,l}(b) = \inf \left\{ t \geq l : \log \hat{\Lambda}_n(t-l+1, t, \hat{\theta}_n) \geq b, \right. \\ \left. \sup_{\|\theta - \hat{\theta}_n\| < 1/\sqrt{b}} \lambda_{\max} \left(-\nabla^2 \log \hat{\Lambda}_n(t-l+1, t, \theta) \right) \leq b, \right. \\ \left. \text{and } \hat{\theta}_n = \arg \max_{\theta} \log \hat{\Lambda}_n(t-l+1, t, \theta) \in \text{Int}(\Theta) \right\}. \end{aligned} \quad (4.46)$$

We define the generalized W-SGLR stopping time as

$$\hat{\tau}_{\text{W-SGLR}}(b) = \min\{\hat{\tau}(b), \hat{\tau}_n(b)\}. \quad (4.47)$$

Note that $\hat{\tau}_{\text{W-SGLR}}$ is a modification of $\tau_{\text{W-SGLR}}$ with additional conditions required for stopping.

The paper [118] first introduced window-limited generalized detection rules. We compute the false alarm probability of $\hat{\eta}_l$ and $\hat{\eta}_{n,l}$. We then use this false alarm probability to obtain a lower bound for the ARL of $\hat{\tau}$ and $\hat{\tau}_n$ in Proposition 4.4.

Lemma 4.4. Suppose that Assumption 4.3 holds. Given any $0 < \delta < 1$, there exists $b_\delta > 0$ such that $\mathbb{P}_{\nu_n, \infty}(\hat{\eta}_k < \infty) \leq \exp(-(1-\delta)b)$ and $\mathbb{P}_{\nu_n, \infty}(\hat{\eta}_{n,l} < \infty) \leq \exp(-(1-\delta)b)$ for any $\nu_n \in \mathbb{N} \cup \{\infty\}$ and $b \geq b_\delta$.

Proof. As the proof is similar to Lemma 2 in [9], we omit it here and refer the reader to the extended version in [123]. $\square$

Proposition 4.4. Suppose that Assumption 4.3 holds. For any $0 < \delta < 1$, there exists $b_\delta > 0$ such that for all $b \geq b_\delta$ and $\nu_n \in \mathbb{N} \cup \{\infty\}$, we have

$$\text{ARL}(\widehat{\tau}_{\text{W-SGLR}}(b)) = \inf_{\nu_n} \mathbb{E}_{\nu_n, \infty} [\widehat{\tau}_{\text{W-SGLR}}(b)] \geq \frac{1}{2} e^{(1-\delta)b}. \quad (4.48)$$

Proof. Fix $0 < \delta < 1$. By Lemma 4.4, there exists $b_\delta \geq 0$ such that $\mathbb{P}_{\nu_n, \infty}(\min\{\widehat{\eta}_1, \widehat{\eta}_{n,1}\} < \infty) \leq 2 \exp(-(1-\delta)b)$ for all $b \geq b_\delta$. Applying results from Theorem 6.16 in [18], we obtain

$$\mathbb{E}_{\nu_n, \infty} [\widehat{\tau}_{\text{W-SGLR}}] \geq \frac{1}{2} e^{(1-\delta)b} \quad (4.49)$$

for all $b \geq b_\delta$. Taking infimum over ν_n , we have $\text{ARL}(\widehat{\tau}_{\text{W-SGLR}}(b)) \geq \frac{1}{2} e^{(1-\delta)b}$, and the proof is complete. $\square$

4.4 Numerical Results

In this section, we first illustrate the performance of the proposed W-SGLR stopping time under the assumption that the distributions $\mathbb{P}^f, \mathbb{P}^{f_n}, \mathbb{P}^g$ and $\mathbb{P}^{g_n}$ are known. Next, we illustrate the performance of the proposed generalised W-SGLR stopping time when $\mathbb{P}^g$ and $\mathbb{P}^{g_n}$ belongs to a parametrized family of distributions. Finally, we evaluate the performance of the proposed W-SGLR stopping time on real data from the Case Western Reserve University Bearing Dataset [35].

4.4.1 W-SGLR on Synthetic Data Satisfying Assumption 4.2

In our first set of simulations, we let $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, and $g_n = \mathcal{N}(2, 10)$ where the critical change is a change in variance and the nuisance change is a change in mean (see example after Assumption 4.2 for motivation). We ran the simulations with two change-point configurations to illustrate the behaviour of the W-SGLR test statistic for different window-sizes. In Fig. 4.2a, we set $\nu_c =$

(A) $\nu_c = 1000, \nu_n = 1500$ (B) $\nu_c = 1500, \nu_n = 1000$

FIG. 4.2: W-SGLR test statistics $S_{W-SGLR}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(2, 10)$ and $I = 3.34$.

1000, $\nu_n = 1500$, while in Fig. 4.2b, we set $\nu_c = 1500, \nu_n = 1000$. In Fig. 4.2a and Fig. 4.2b, the test statistic $S_{W-SGLR}(t)$ remains low before the critical change-point and grows linearly with the gradient of at least $I = 3.34$, as described in (4.19), after the critical change-point. This trend continues until the test statistic approximately achieves the value of $m_b I$. From our choice of m_b in (4.18), we see that $m_b I > b$ for b large, i.e., our τ_{W-SGLR} is able to detect the critical change given sufficient delay for every choice of b sufficiently large. However, choosing a larger m_b is more resistant to outlier noise. For example, in Fig. 4.2a, when $m_b = 1024$, we note that the test statistic continues to grow linearly with the gradient I even after the nuisance change point. The trade-off is the increase in memory requirement and computational complexity. In Fig. 4.2b, we note that the test statistic continues to remain low during the period between the nuisance and the critical change point. This demonstrates that τ_{W-SGLR} is oblivious to the nuisance change.

FIG. 4.3: Comparison of the trade-off performance for τ_{W-SGLR} , τ_{GLRT} , τ_{FMA} and $\tau_{2-stage}$.

Next, we compare $\tau_{\text{W-SGLR}}$, the GLRT stopping time τ_{GLRT} developed in [124], the finite-moving average (FMA) stopping time τ_{FMA} and a naive 2-stage CuSum stopping time denoted as $\tau_{2\text{-stage}}$. Following ideas from the TCD literature [117], the FMA stopping time τ_{FMA} is constructed by replacing the maximum in the test statistic (4.16) with a sum across the entire window, i.e. setting $k = t - m_b$. It should be noted that while the FMA stopping time has been shown to perform well for the TCD problem, there are no guarantees that it will perform as well for the QCD problem. The naive stopping time $\tau_{2\text{-stage}}$ is constructed from stopping times based on the CuSum stopping time described in (4.6) with $\tau_{\{p \rightarrow q\}}(b) = \inf \left\{ t : \max_{1 \leq k \leq t+1} \prod_{i=k}^t \frac{p(x_i)}{q(x_i)} > e^b \right\}$ for any pair of pdfs p with $q \neq p$. We consider four stopping times: $\tau_{f \rightarrow f_n}(b_n)$, $\tau_{f \rightarrow g}(b_c)$, $\tau_{f \rightarrow g_n}(b_c)$, and $\tau_{f_n \rightarrow g_n}(b_c)$, where the threshold for declaring a critical change is b_c and the threshold for declaring a nuisance change is b_n . In the first stage, we apply the stopping times $\tau_{f \rightarrow g}(b_c)$, $\tau_{f \rightarrow g_n}(b_c)$ and $\tau_{f \rightarrow f_n}(b_n)$ to the observations. If $\tau_{f \rightarrow g}(b_c)$ or $\tau_{f \rightarrow g_n}(b_c)$ stops the process before $\tau_{f \rightarrow f_n}(b_n)$, we declare that a critical change has occurred and set $\tau_{2\text{-stage}} = \min\{\tau_{f \rightarrow g}(b_c), \tau_{f \rightarrow g_n}(b_c)\}$. Otherwise, we apply $\tau_{f_n \rightarrow g_n}(b_c)$ to the rest of the observations after the stopping time $\tau_{f \rightarrow f_n}(b_n)$ and set $\tau_{2\text{-stage}} = \tau_{f_n \rightarrow g_n}(b_c)$.

In our simulations, our signal is generated using $f = \mathcal{N}(0, 1)$, $g = \mathcal{N}(0.5, 1)$, $f_n = \mathcal{N}(0, 2)$, $g_n = \mathcal{N}(0.5, 2)$. Here, the critical change is a change in mean from 0 to 0.5, and the nuisance change is a change in variance from 1 to 2. We generate a signal of length $2^{16} = 65536$ and independently select the nuisance change point and critical change point with uniform probability on the 2^{16} possible data points. A total of $2^{12} = 4096$ signals are generated. We compare the trade-off between the ADD and the empirical ARL of the proposed $\tau_{\text{W-SGLR}}$, τ_{GLR} , τ_{FMA} and $\tau_{2\text{-stage}}$ in Fig. 4.3. We observe that our proposed $\tau_{\text{W-SGLR}}$ achieves a lower ADD as compared to τ_{FMA} , τ_{GLR} and $\tau_{2\text{-stage}}$ for large empirical ARL.

In the next set of simulations, we let $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, and $g_n = \mathcal{N}(5, 20)$ where unlike the first set of simulations, the change in mean before and after the critical change point differs, and the change in variance before and after the nuisance change point differs. In Fig. 4.4a, we set $\nu_c = 1000 < \nu_n = 1500$.

(A) $\nu_c = 1000, \nu_n = 1500$ (B) $\nu_c = 1500, \nu_n = 1000$

FIG. 4.4: W-SGLR test statistics $S_{W-SGLR}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(5, 20)$ and $I = 3.34$.

We see that the test statistic $S_{\text{W-SGLR}}(t)$ remains low before the critical change-point and grows linearly with a gradient of at least $I = 3.34$ (cf. (4.19)), after the critical change-point. This trend continues until the nuisance change point ν_n where rate of growth changes to $\mathbb{E}^g \left[\log \frac{g_n}{f_n} \right] > I$ until it approximately achieves the value of $m_b \mathbb{E}^g \left[\log \frac{g_n}{f_n} \right]$. While the growth of the test statistic after the change-point is not linear, the observation that the overall rate of growth from the critical change point is at least I is consistent with Lemma 4.2. In Fig. 4.4b, we set $\nu_c = 1500 > \nu_n = 1000$ and note that the test statistic continues to remain low during the period between the nuisance and the critical change points. This demonstrates that $\tau_{\text{W-SGLR}}$ is oblivious to the nuisance change prior to the critical change-point.

4.4.2 W-SGLR on Synthetic Data Violating Assumption 4.2

When Assumption 4.2 is violated, Theorem 4.3 still provides the asymptotic trade-off between the ARL and the WADD. However, the asymptotic optimality of the W-SGLR stopping is not guaranteed. Here, we provide discussions and numerical simulations that suggests that the W-SGLR stopping time out-performs the two-stage stopping time and FMA stopping time with respect to (4.2).

If Assumption 4.2 is violated, from Theorem 4.3, we have

$$\text{WADD}(\tau_{\text{W-SGLR}}(b)) \leq (D(g \| f_n)^{-1} + o(1)) b.$$

This worst-case performance of the W-SGLR stopping time is achieved when $\nu_n = \infty$. For the rest of this discussion, we let $\nu_n = \infty$ to compare the two-stage stopping time with our proposed W-SGLR stopping time under this worst-case scenario.

Since $D(g \| f) \geq D(g \| f_n)$, the CuSum

$$\max_{1 \leq k \leq t+1} \sum_{i=k}^t \log \frac{f_n(X_i)}{f(X_i)}$$

FIG. 4.5: Comparison of the trade-off performance for $\tau_{\text{W-SGLR}}$, $\tau_{2\text{-stage}}$ and τ_{FMA} when Assumption 4.2 is violated with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 5)$, $g = \mathcal{N}(3, 10)$ and $g_n = \mathcal{N}(5, 10)$.

associated with the stopping time $\tau_{f \rightarrow f_n}$ experiences a positive drift when $\nu_c \leq t < \nu_n$. Thus, for any finite threshold b_n and sufficiently large b_c , the two-stage stopping time declares that a nuisance change has taken place and transits into the second stage after the critical change point. The CuSum associated with the stopping time $\tau_{f_n \rightarrow g_n}$ in the second stage is expected to grow at a rate of $\mathbb{E}^g \left[\log \frac{g_n}{f_n} \right]$ for when $\nu_c \leq t < \nu_n$. In contrast, the W-SGLR test statistic, from Theorem 4.2, is expected to grow at a rate of $\mathbb{E}^g \left[\log \frac{g}{f_n} \right] \geq \mathbb{E}^g \left[\log \frac{g_n}{f_n} \right]$. Heuristically, this means that, when $\nu_n = \infty$, the $\text{WADD}(\tau_{2\text{-stage}}) \geq \text{WADD}(\tau_{\text{W-SGLR}})$ as the $\text{ARL} \rightarrow \infty$. It should also be noted that it is possible that the stopping time $\tau_{2\text{-stage}}$ fails completely when $\mathbb{E}^g \left[\log \frac{g_n}{f_n} \right]$ is negative and $\nu_n = \infty$.

In Figs. 4.5a and 4.5b, we compare the trade-off between the ADD and the ARL of the different stopping times when Assumption 4.2 is violated under the cases $(\nu_n, \nu_c) = (2000, 4000)$ and $(\nu_n, \nu_c) = (4000, 2000)$, respectively. To estimate the empirical ARL, the stopping times are applied to a set of 4096 signals each of length 2^{16} with nuisance change point independently selected with uniform probability on the 2^{16} possible data points. To compute the corresponding ADD for the stopping times, they are applied to a set of 4096 signals of length 4500. It can be seen from both Figs. 4.5a and 4.5b that the W-SGLR stopping time achieves a lower ADD as compared to both $\tau_{2\text{-stage}}$ and τ_{FMA} for large empirical ARL. Consistent with our intuition, it can be seen that $\tau_{\text{W-SGLR}}$ significantly outperforms $\tau_{2\text{-stage}}$ when $\nu_c < \nu_n$.

4.4.3 Parametrized Post-Change Distributions

In this set of simulations, we let $f = \mathcal{N}(0, 1)$, the critical change to be a change in variance where $g = \mathcal{N}(0, \theta^2)$, and the nuisance change to be a change in the mean where $f_n = \mathcal{N}(2, 1)$ and $g_n = \mathcal{N}(2, \theta_n^2)$. The parameters $\theta = \theta_n = 10$ are unknown to the change detection algorithms. This corresponds to the case where the transmission power is unknown in the problem of spectrum sensing [119]. We ran the simulations with two change-point configurations to demonstrate the

(A) $\nu_c = 333$, $\nu_n = 666$ (B) $\nu_c = 666$, $\nu_n = 333$

FIG. 4.6: The generalized W-SGLR test statistics $\hat{S}_{W-SGLR}(t)$ with $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(2, 1)$, $g = \mathcal{N}(0, 10)$, $g_n = \mathcal{N}(2, 10)$ and $I = 3.34$.

behavior of the generalized W-SGLR test statistic used in $\hat{\tau}_{\text{W-SGLR}}$ as described in (4.42) and (4.45) for window-sizes $m_b = 16, 32, 64, 128, 256, 512$ and 1024. In Fig. 4.6a, we set the critical change point to be $\nu_c = 333$ and nuisance change point to be $\nu_n = 666$. It can be observed that our proposed generalized W-SGLR test statistic remains low during the pre-change regime, increases in the post-change regime and continue to increase in the nuisance post-change regime when the window is sufficiently large. This demonstrates that our stopping time $\hat{\tau}_{\text{W-SGLR}}$ is able to detect the critical change even in the nuisance critical change region. In Fig. 4.6b, we set the critical change point to be $\nu_c = 666$ and nuisance change point to be $\nu_n = 333$.

We see that our stopping time $\hat{\tau}_{\text{W-SGLR}}$ is effective in detecting critical changes while ignoring the nuisance change in pre-change regime for window sizes as small as $m_b = 16$. In practice, we can use graphs like Fig. 4.6a and Fig. 4.6b to compare if the increase in the test-statistic after the critical change is discernible from the test-statistic in the pre-change regime. This would provide assistance in determining if the choice window-size is suitable.

Next, we compare the generalized W-SGLR stopping time with the W-SGLR stopping time. In our simulations, our signal is generated using $f = \mathcal{N}(0, 1)$, $f_n = \mathcal{N}(0, 2)$, $g = \mathcal{N}(\theta, 1)$, $g_n = \mathcal{N}(\theta_n, 2)$. Here we set $\theta = \theta_n = 2$ and assume that the condition that $\hat{\theta} \in \text{Int}(\Theta)$ is always satisfied. We generate a signal of length $2^{16} = 65,536$ and independently select the nuisance change point and critical change point with uniform probability on the 2^{16} possible data points. A total of $2^{12} = 4096$ signals are generated. We compare the trade-off between the ADD and the ARL of the proposed W-SGLR stopping time when θ and θ_n are known against the generalized W-SGLR stopping time when θ and θ_n are unknown in Fig. 4.7. We observe that the generalized W-SGLR stopping time has a higher ADD as compared to the W-SGLR stopping time. Our experiments suggest that the difference in ADD is bounded as the ARL becomes large.

FIG. 4.7: Comparison of trade-off performance for the proposed stopping time $\hat{\tau}_{W-SGLR}$ and τ_{W-SGLR} .

4.4.4 Real Data

In this subsection, we test our proposed stopping time τ_{W-SGLR} on the Case Western Reserve University Bearing Dataset [35]. The dataset is collected from experiments conducted using an electric motor with accelerometer data measured at locations near to and remote from the motor bearings. Samples were collected at 12 KHz. We pre-process the signal by de-trending the signal using a first order finite difference: for each signal sample time t , let $X_t = Y_t - Y_{t-1}$, where Y_t is the observed raw signal sample at time t .

We consider signals X_t obtained at a motor load of 1hp and 2hp with normal bearings and also faulty bearings with a 0.007-inch fault diameter. We assume that the critical change would be the transition from a normal to faulty bearing, and a nuisance change would be a change in the motor load. We use the first 12,000 samples as training data to build a model for each of the following scenarios: normal bearings under a motor load of 1hp, normal bearings under a motor load of 2hp,

FIG. 4.8: Plots of the pdfs of the observed de-trended signal in different scenarios.

faulty bearings under a motor load of 1hp, and faulty bearings under a motor load of 2hp. Fig. 4.8 shows the learned pdf of the de-trended signals observed in each scenario.

There are two challenges faced in testing our proposed stopping time on real data: (i) we lack theoretical results for the ARL of the 2-stage stopping times for the selection of appropriate thresholds for comparison and (ii) real run-to-failure data is difficult to obtain. We divide the remaining samples into 3 disjoint sets to address the above challenges.

For the first set, we create a training set of 1000 signals each with length 36,000 with a randomly selected nuisance change point ν_n for each signal such that there is a period of $\nu_n - 1$ samples for a normal bearing under a motor load of 1hp, and a period of $36,000 - \nu_n + 1$ samples for a normal bearing under a motor load of 2hp. We select appropriate thresholds for each of the stopping times so that the empirical ARL varies between 1200 and 18,000.

FIG. 4.9: Examples of the W-SGLR test statistic with $m_b = 1024$.

The next two sets are testing sets. We create 1000 signals of length 3600 each with (i) a period of 1200 samples for a normal bearing under a motor load of 1hp, which transitions to (ii) a period of 1200 samples for a normal bearing under a motor load of 2hp, which finally transitions to (iii) a period of 1200 samples for a faulty bearing under a motor load of 2hp.

Similarly, we create 1000 signals for the scenario where a normal bearing under a motor load of 1hp transitions to a faulty bearing under a motor load of 1hp and finally a faulty bearing under a motor load of 2hp.

Finally, we apply the selected thresholds obtained from the first training set to the two testing sets to compute the stopping times' empirical ADD performance. The window size of $m_b = 8$ for the FMA stopping time is selected to minimize its empirical ADD on the test set. For this dataset, if m_b is chosen to be 128 or 1024, the empirical ADD of the FMA stopping time becomes much larger compared to the empirical ADD of the W-SGLR and 2-stage stopping times. Thus, we only present the ARL-ADD trade-off for $m_b = 8$.

In Figs. 4.9a and 4.9b, we present some examples of the performance of the W-SGLR test statistic. It can be seen that in both cases, the test-statistic remains low before the bearing failure and quickly rises after the bearing fails even as the motor load changes.

In Fig. 4.10a and Fig. 4.10b, we present the trade-off between the empirical ADD and ARL for the proposed W-SGLR stopping time with $m_b = 1024$, the 2-stage stopping times with different thresholds b_n and the FMA stopping time. It can be seen that our proposed stopping time $\tau_{\text{W-SGLR}}$ achieves better ADD-ARL trade-off compared to the other stopping times. However, as the KL divergences

$D(g_n \| f_n), D(g \| f_n), D(g_n \| f), D(g \| f)$ are large, the empirical ADD for all the algorithms remains low across the range of ADD tested. In this case, the reduction in empirical ADD is small, between 1 to 4 samples, over the range of ARLs tested. In terms of computational complexity, up till sample t , the W-SGLR stopping time requires $O(m_b t)$ operations [125], which is slightly more than both the two-stage

stopping time and the FMA stopping time, both of which require $O(t)$ operations. Thus, for applications that have limited computational resources and large differences in their pre and post-change distributions, we may want to consider using the FMA or the 2-stage stopping time as the degradation in performance is small.

Appendix 4.A Proof of Propositions 4.1 and 4.2

We start off with some notation definitions. Let $L_i = \log \frac{f_n(X_i)}{f(X_i)}$. For any $N \geq 0$, let $L_{i,>N} = L_i \mathbf{1}_{\{|L_i|>N\}}$ and $L_{i,\leq N} = L_i \mathbf{1}_{\{|L_i|\leq N\}}$. For any $k, t \in \mathbb{N}$ such that $k \leq t$, we define the following averages:

$$\begin{aligned} \overline{L^{k:t}} &= \frac{1}{t-k+1} \sum_{i=k}^t L_i, \quad \overline{L_{>N}^{k:t}} = \frac{1}{t-k+1} \sum_{i=k}^t L_{i,>N}, \\ \overline{L_{\leq N}^{k:t}} &= \frac{1}{t-k+1} \sum_{i=k}^t L_{i,\leq N} \end{aligned} \quad (4.50)$$

We have

$$\overline{L^{k:t}} = \overline{L_{>N}^{k:t}} + \overline{L_{\leq N}^{k:t}}. \quad (4.51)$$

For the case where $k > t$, we let $\overline{L^{k:t}} = \overline{L_{>N}^{k:t}} = \overline{L_{\leq N}^{k:t}} = 0$. Finally, we define the random variable

$$V_{k,t} = \arg \max_{k \leq j \leq t+1} \prod_{i=k}^{j-1} f(X_i) \prod_{i=j}^t f_n(X_i). \quad (4.52)$$

An outline of the proof of Propositions 4.1 and 4.2 is as follows. Lemma 4.A.2 and Lemma 4.A.4 provide the results required for controlling the error bound in Proposition 4.1. Similarly, Lemma 4.A.3 and Lemma 4.A.4 provide the results required for controlling the error bound in Proposition 4.2. The Lemmas 4.A.2, 4.A.3 and 4.A.4 require that decay in the tail probabilities of the average log-likelihood ratio $\frac{1}{n} \sum_{i=k}^{k+n-1} \log \frac{f_n(X_i)}{f(X_i)}$ to be at most $O(n^{-2})$, which is shown in Lemma 4.A.1.

FIG. 4.10: Trade-off between the empirical ADD and ARL.

Lemma 4.A.1. For any $\nu_c, \nu_n, k, l, n \in \mathbb{N}$ such that $\nu_c \leq k$ and $\max\{\nu_n, \nu_c\} \leq l$, and $\epsilon > 0$, we have

$$\mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L^{k:k+n-1}} - \rho_g \right| \geq \epsilon \right) \leq \frac{K_g}{\epsilon^4 n^2}, \quad (4.53)$$

$$\mathbb{P}_{\nu_n, \nu_c} \left(\left| \overline{L^{l:l+n-1}} - \rho_{g_n} \right| \geq \epsilon \right) \leq \frac{K_{g_n}}{\epsilon^4 n^2}, \quad (4.54)$$

where $K_g = \omega_g^4 + \frac{3}{2}\sigma_g^4$ and $K_{g_n} = \omega_{g_n}^4 + \frac{3}{2}\sigma_{g_n}^4$.

Proof. As the proof is elementary, we omit it here and refer the reader to the extended version in [123]. $\square$

From Lemma 4.A.1, for any $\nu_c \leq k \leq v \leq t < \infty$, we have for $\rho_g < 0$,

$$\begin{aligned} & \mathbb{P}_{\infty, \nu_c} (V_{k,t} = v) \\ & \leq \mathbb{P}_{\infty, \nu_c} (|\overline{L^{v:t}} - \rho_g| \geq |\rho_g|) \\ & \leq \frac{K_g}{|\rho_g|^4 (t - v + 1)^2}. \end{aligned} \quad (4.55)$$

Similarly, for $\max\{\nu_c, \nu_n\} \leq k < v \leq t + 1 < \infty$, and $\rho_{g_n} > 0$, we have

$$\mathbb{P}_{\nu_n, \nu_c} (V_{k,t} = v) \leq \frac{K_g}{|\rho_{g_n}|^4 (v - k)^2}. \quad (4.56)$$

For the next two lemmas, we use bounds on the tail probability of $\overline{L_{>N_g}^{k:t}}$ to derive asymptotic properties of the random variable $V_{k,t}$ under the distributions $\mathbb{P}_{\nu_n, \nu_c}$ and $\mathbb{P}_{\infty, \nu_c}$ for any $\nu_n, \nu_c \in \mathbb{N}$.

Lemma 4.A.2. For any $0 < c < 1$, $\nu_c \leq k < \infty$, and $\rho_g < 0$, we have

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} > c \right) = 0. \quad (4.57)$$

Proof. We have

$$\begin{aligned}
& \mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} > c \right) \\
& \leq \sum_{v=k}^{\lfloor (1-c)(t+1) + ck \rfloor} \mathbb{P}_{\infty, \nu_c} (V_{k,t} = v) \\
& \leq \sum_{v=k}^{\lfloor (1-c)(t+1) + ck \rfloor} \frac{K_g}{|\rho_g|^4 (t - v + 1)^2} \\
& \leq (\lfloor (1-c)(t+1) + ck \rfloor - k + 1) \frac{K_g}{|\rho_g|^4 (t - k + 1)^2} \rightarrow 0,
\end{aligned} \tag{4.58}$$

as $t \rightarrow \infty$. The inequality (4.58) follows from (4.55). The proof is now complete. $\square$

Lemma 4.A.3. For any $0 < c < 1$ and $\nu_c \leq k < \infty$, let $k' = \max\{k, \nu_n\}$. If $\rho_{g_n} > 0$, we have

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\nu_n, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} > c \right) = 0. \tag{4.59}$$

Proof. For $t > \frac{k-k'}{c} + k - 1$, we have $c(t - k + 1) + k' > k$ and

$$\begin{aligned}
& \mathbb{P}_{\nu_n, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} > c \right) \\
& = \sum_{v=\lceil c(t-k+1) + k' \rceil}^{t+1} \mathbb{P}_{\nu_n, \nu_c} (V_{k,t} = v) \\
& \leq \sum_{v=\lceil c(t-k+1) + k' \rceil}^{t+1} \frac{K_{g_n}}{|\rho_{g_n}|^4 (v - k)^2} \\
& \leq (t + 1 - \lceil c(t - k + 1) + k' \rceil + 1) \frac{K_{g_n}}{|\rho_{g_n}|^4 (\lceil c(t - k + 1) \rceil + k' - k)^2} \rightarrow 0,
\end{aligned} \tag{4.60}$$

as $t \rightarrow \infty$, where (4.60) follows from (4.56). The proof is now complete. $\square$

Lemma 4.A.4. Suppose that $\rho_g < 0$ and $\rho_{g_n} > 0$. For any $\epsilon, \delta > 0$ and $k \geq \max\{\nu_n, \nu_c\}$, there exist $N_{g_n}, N_g \in \mathbb{N}$ such that for any $t \geq k$,

$$\mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L_{>N_g}^{V_{k,t}, t}} \right| \geq \epsilon \right) \leq \delta, \tag{4.61}$$

$$\mathbb{P}_{\nu_n, \nu_c} \left(\left| \overline{L_{>N_{g_n}}^{k:V_{k,t}-1}} \right| \geq \epsilon \right) \leq \delta. \quad (4.62)$$

Proof. Given any $\epsilon, \delta > 0$, since the fourth moment of $\log \frac{f_n(X)}{f(X)}$ exists, by the monotone convergence theorem, there exists N_g and N_{g_n} such that

$$\mathbb{E}^{g_n} \left[\left| \log \frac{f_n(X)}{f(X)} \right|^4 \mathbf{1}_{\left\{ \left| \log \frac{f_n(X)}{f(X)} \right| > N_{g_n} \right\}} \right] \leq \frac{\epsilon^4 \delta^4}{M^4}, \quad (4.63)$$

$$\mathbb{E}^g \left[\left| \log \frac{f_n(X)}{f(X)} \right|^4 \mathbf{1}_{\left\{ \left| \log \frac{f_n(X)}{f(X)} \right| > N_g \right\}} \right] \leq \frac{\epsilon^4 \delta^4}{M^4}, \quad (4.64)$$

where $M = \sum_{v=1}^{\infty} \left(\frac{K_g}{|\rho_g|^4 v^2} \right)^{\frac{3}{4}} < \infty$. Applying Markov's inequality, we obtain

$$\mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \geq \epsilon \right) \leq \frac{1}{\epsilon} \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \right]. \quad (4.65)$$

Next, we derive an upper bound for $\mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \right]$. For any $v \leq t$, we have

$$\begin{aligned} \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{v:t}} \right|^4 \right] &= \mathbb{E}_{\infty, \nu_c} \left[\left| \frac{1}{t-v+1} \sum_{i=v}^t L_{i, >N_g} \right|^4 \right] \\ &\leq \frac{1}{t-v+1} \sum_{i=v}^t \mathbb{E}_{\infty, \nu_c} \left[\left| L_{i, >N_g} \right|^4 \right] \end{aligned} \quad (4.66)$$

$$\leq \frac{\epsilon^4 \delta^4}{M^4}, \quad (4.67)$$

where (4.66) follows from Jensen's inequality, and (4.67) follows from (4.63). We obtain

$$\begin{aligned} \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \right] &= \sum_{v=k}^{t+1} \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \mathbf{1}_{\{V_{k,t}=v\}} \right] \\ &= \sum_{v=k}^t \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{v:t}} \right| \mathbf{1}_{\{V_{k,t}=v\}} \right] \end{aligned} \quad (4.68)$$

$$\leq \sum_{v=k}^t \mathbb{E}_{\infty, \nu_c} \left[\left| \overline{L_{>N_g}^{v:t}} \right|^4 \right]^{\frac{1}{4}} \mathbb{P}_{\infty, \nu_c} (V_{k,t} = v)^{\frac{3}{4}} \quad (4.69)$$

$$\leq \frac{\epsilon \delta}{M} \sum_{v=k}^t (\mathbb{P}_{\infty, \nu_c} (V_{k,t} = v))^{\frac{3}{4}}, \quad (4.70)$$

$$\leq \frac{\epsilon\delta}{M} \sum_{v=k}^t \left(\frac{K_g}{|\rho_g|^4(t-v+1)^2} \right)^{\frac{3}{4}} \quad (4.71)$$

$$\leq \epsilon\delta. \quad (4.72)$$

where (4.68) is because $\overline{L_{>N_g}^{t+1:t}} = 0$, (4.69) follows from Hölder's inequality, (4.70) from (4.67), (4.71) from (4.55), and (4.72) from the definition of M . From (4.65), we have $\mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L_{>N_g}^{V_{k,t}:t}} \right| \geq \epsilon \right) \leq \frac{\epsilon\delta}{\epsilon} = \delta$, and (4.61) is proved. The proof of (4.62) is similar and the lemma is proved. $\square$

4.A.1 Proof of Proposition 4.1

It suffices to show that for any $\epsilon, \delta > 0$, there exists T such that for all $t \geq T$ we have

$$\mathbb{P}_{\infty, \nu_c} \left(\left| \frac{\log \Lambda(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g(X_i)}{f(X_i)} \right| \geq \epsilon \right) \leq \delta. \quad (4.73)$$

For any $N \geq 0$ and $c > 0$, the left-hand side of (4.73) becomes

$$\begin{aligned} & \mathbb{P}_{\infty, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=V_{k,t}}^t \log L_i \right| \geq \epsilon \right) \\ & \leq \mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} \left| \overline{L_{>N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right) \\ & \quad + \mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} \left| \overline{L_{\leq N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right) \\ & \leq \mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L_{>N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right) \\ & \quad + \mathbb{P}_{\infty, \nu_c} \left(\left\{ \frac{t - V_{k,t} + 1}{t - k + 1} \left| \overline{L_{\leq N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right\} \cap \left\{ \frac{t - V_{k,t} + 1}{t - k + 1} \leq c \right\} \right) \\ & \quad + \mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} > c \right). \end{aligned} \quad (4.74)$$

From Lemma 4.A.4, there exists N such that $\mathbb{P}_{\infty, \nu_c} \left(\left| \overline{L_{>N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right) \geq \frac{\delta}{2}$. Next, by choosing $c = \frac{\epsilon}{4N}$, we have $\mathbb{P}_{\infty, \nu_c} \left(\left\{ \frac{t - V_{k,t} + 1}{t - k + 1} \left| \overline{L_{\leq N}^{V_{k,t}:t}} \right| \geq \frac{\epsilon}{2} \right\} \cap \left\{ \frac{t - V_{k,t} + 1}{t - k + 1} \leq c \right\} \right) = 0$. Finally, from Lemma 4.A.2, there exists T such that for all $t \geq T$, we have

$\mathbb{P}_{\infty, \nu_c} \left(\frac{t - V_{k,t} + 1}{t - k + 1} > c \right) \leq \frac{\delta}{2}$. The right-hand side of (4.74) is then upper bounded by δ , and the proof is complete.

4.A.2 Proof of Proposition 4.2

It suffices to show that for any $\epsilon, \delta > 0$, there exists T such that for all $t \geq T$ we have

$$\mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{\log \Lambda(k, t)}{t - k + 1} - \frac{1}{t - k + 1} \sum_{i=k}^t \log \frac{g(X_i)}{f_n(X_i)} \right| \geq \epsilon \right) \leq \delta. \quad (4.75)$$

Let $k' = \max\{k, \nu_n\}$. The left-hand side of (4.75) can be written as

$$\begin{aligned} & \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=k}^{V_{k,t}-1} \log L_i \right| \geq \epsilon \right) \\ & \leq \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=k}^{k'-1} \log L_i \right| \geq \frac{\epsilon}{2} \right) \end{aligned} \quad (4.76)$$

$$+ \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=k'}^{V_{k,t}-1} \log L_i \right| \geq \frac{\epsilon}{2} \right). \quad (4.77)$$

Applying Markov's inequality to (4.76), there exists T_1 such that for all $t \geq T_1$, we have

$$\begin{aligned} & \mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=k}^{k'-1} \log L_i \right| \geq \frac{\epsilon}{2} \right) \\ & \leq \frac{2}{\epsilon(t - k + 1)} \mathbb{E}_{\nu_n, \nu_c} \left[\left| \sum_{i=k}^{k'-1} \log L_i \right| \right] < \frac{\delta}{3}. \end{aligned}$$

For any $N \geq 0$ and $c \geq 0$, (4.77) becomes

$$\mathbb{P}_{\nu_n, \nu_c} \left(\left| \frac{1}{t - k + 1} \sum_{i=k'}^{V_{k,t}-1} \log L_i \right| \geq \frac{\epsilon}{2} \right)$$

$$\begin{aligned}
&\leq \mathbb{P}_{\nu_n, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} \left| \overline{L_{>N}^{k':V_{k,t}-1}} \right| \geq \frac{\epsilon}{4} \right) \\
&\quad + \mathbb{P}_{\nu_n, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} \left| \overline{L_{\leq N}^{k':V_{k,t}-1}} \right| \geq \frac{\epsilon}{4} \right) \\
&\leq \mathbb{P}_{\nu_n, \nu_c} \left(\left| \overline{L_{>N}^{k':V_{k,t}}} \right| \geq \frac{\epsilon}{4} \right) \\
&\quad + \mathbb{P}_{\nu_n, \nu_c} \left(\left\{ \frac{V_{k,t} - k'}{t - k + 1} \left| \overline{L_{\leq N}^{k':V_{k,t}}} \right| \geq \frac{\epsilon}{2} \right\} \cap \left\{ \frac{V_{k,t} - k'}{t - k + 1} \leq c \right\} \right) \\
&\quad + \mathbb{P}_{\nu_n, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} > c \right). \tag{4.78}
\end{aligned}$$

From Lemma 4.A.4, there exists N such that $\mathbb{P}_{\nu_n, \nu_c} \left(\left| \overline{L_{>N}^{k':V_{k,t}-1}} \right| > \frac{\epsilon}{4} \right) < \frac{\delta}{3}$. Next, by choosing $c = \frac{\epsilon}{4N}$, we have $\mathbb{P}_{\nu_n, \nu_c} \left(\left\{ \frac{V_{k,t} - k'}{t - k + 1} \left| \overline{L_{\leq N}^{k':V_{k,t}-1}} \right| > \frac{\epsilon}{4} \right\} \cap \left\{ \frac{V_{k,t} - k'}{t - k + 1} \leq c \right\} \right) = 0$. Finally, from Lemma 4.A.2, there exists T_2 such that for all $t \geq T_2$, we have $\mathbb{P}_{\infty, \nu_c} \left(\frac{V_{k,t} - k'}{t - k + 1} > c \right) < \frac{\delta}{3}$. The right-hand side of (4.78) is then upper bounded by δ , and the proof is complete.

Appendix 4.B Proof of Lemma 4.1

For any $b > 0$, we have

$$\begin{aligned}
&\mathbb{P}_{\nu_n, \infty} (\eta^1 < \infty) \\
&= \sum_{k=1}^{\infty} \mathbb{P}_{\nu_n, \infty} (\eta^1 = k) \\
&= \sum_{k=1}^{\infty} \int_{\{\eta^1=k\}} \prod_{i=1}^k h_{\nu_n, \infty, i}(x_i) \, d\mathbf{x}_{1:k} \\
&\leq \sum_{k=1}^{\infty} \int_{\{\eta^1=k\}} e^{-b} \Lambda(1, k) \prod_{i=1}^k h_{\nu_n, \infty, i}(x_i) \, d\mathbf{x}_{1:k} \\
&= e^{-b} \sum_{k=1}^{\infty} \int_{\{\eta^1=k\}} \frac{\prod_{i=1}^k g(x_i)}{\max_{1 \leq j \leq k+1} \prod_{i=1}^{j-1} f(x_i) \prod_{i=j}^k f_n(x_i)} \\
&\quad \times \prod_{i=1}^k h_{\nu_n, \infty, i}(x_i) \, d\mathbf{x}_{1:k}
\end{aligned}$$

$$\begin{aligned}
&\leq e^{-b} \sum_{k=1}^{\infty} \int_{\{\eta^1=k\}} \prod_{i=1}^k g(x_i) \, d\mathbf{x}_{1:k} \\
&\leq e^{-b} \mathbb{P}_{\infty,1}(\eta^1 < \infty) \leq e^{-b}.
\end{aligned}$$

The proof that $\mathbb{P}_{\nu_n,\infty}(\eta_n^1 < \infty) \leq e^{-b}$ is similar. We then have $\mathbb{P}_{\nu_n,\infty}(\min\{\eta^1, \eta_n^1\} < \infty) \leq 2e^{-b}$. Since $\tau_{\text{SGLR}} = \min\{\tau(b), \tau_n(b)\} = \inf_{k \geq 1} \min\{\eta^k, \eta_n^k\}$, applying [18, Theorem 6.16], we obtain $\mathbb{E}_{\nu_n,\infty}[\tau_{\text{SGLR}}(b)] \geq \frac{1}{2}e^b$ and (4.31) follows from $\tau_{\text{W-SGLR}} \geq \tau_{\text{SGLR}}$. The proof is now complete.

Appendix 4.C Proof of Lemma 4.2

It suffices to show that for any $\epsilon > 0$, there exists T such that for all $t \geq T$ we have

$$\sup_{\nu_n \in \mathbb{N}, 1 \leq \nu_c \leq k} \mathbb{P}_{\nu_n, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k+t-1) - I \leq -\delta \right) \leq \epsilon. \quad (4.79)$$

The set over which the supremum in (4.79) is taken can be divided into two subsets: $A_1 = \{(\nu_n, \nu_c, k) : \max\{\nu_c, \nu_n\} \leq k \leq t\}$ and $A_2 = \{(\nu_n, \nu_c, k) : \nu_c \leq k < \nu_n \leq t\}$. We have

$$\begin{aligned}
&\sup_{A_1} \mathbb{P}_{\nu_n, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k+t-1) - I \leq -\delta \right) \\
&= \mathbb{P}_{1,1} \left(\frac{1}{t} \log \Lambda(1, t) - I \leq -\delta \right) \leq \frac{\epsilon}{2},
\end{aligned} \quad (4.80)$$

where the last inequality follows from Theorem 4.2 for t sufficiently large.

If $\nu_c \leq k < \nu_n \leq t$, we obtain

$$\begin{aligned}
&\mathbb{P}_{\nu_n, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k+t-1) - I \leq -\delta \right) \\
&= \mathbb{P}_{\nu_n-k+1,1} \left(\frac{1}{t} \log \Lambda(1, t) - I \leq -\delta \right) \\
&\leq \mathbb{P}_{\nu_n-k+1,1} \left(\frac{\nu_n-k}{t} \left(\frac{\log \Lambda(1, \nu_n-k)}{\nu_n-k} - I \right) \leq -\frac{\delta}{2} \right) \\
&\quad + \mathbb{P}_{\nu_n-k+1,1} \left(\frac{t-(\nu_n-k)}{t} \left(\frac{\log \Lambda(\nu_n-k+1, t)}{t-(\nu_n-k)} - I \right) \leq -\frac{\delta}{2} \right)
\end{aligned}$$

$$\leq \mathbb{P}_{\infty,1} \left(\frac{\nu_n - k}{t} \left(\frac{\log \Lambda(1, \nu_n - k)}{\nu_n - k} - I \right) \leq -\frac{\delta}{2} \right) \quad (4.81)$$

$$+ \mathbb{P}_{1,1} \left(\frac{t - (\nu_n - k)}{t} \left(\frac{\log \Lambda(1, t - (\nu_n - k))}{t - (\nu_n - k)} - I \right) \leq -\frac{\delta}{2} \right). \quad (4.82)$$

From Theorem 4.2, there exists N_1 such that for all $n \geq N_1$, we have

$$\mathbb{P}_{\infty,1} \left(\frac{1}{n} \log \Lambda(1, n) - I \leq -\frac{\delta}{2} \right) \leq \frac{\epsilon}{2}, \quad (4.83)$$

$$\mathbb{P}_{1,1} \left(\frac{1}{n} \log \Lambda(1, n) - I \leq -\frac{\delta}{2} \right) \leq \frac{\epsilon}{2}. \quad (4.84)$$

From Markov's inequality and Assumption 4.1, there exists N_2 such that for all $1 \leq n < N_1$ and $m \geq N_2$, we have

$$\begin{aligned} \mathbb{P}_{\infty,1} \left(\frac{n}{m} \left(\frac{1}{n} \log \Lambda(1, n) - I \right) \leq -\frac{\delta}{2} \right) \\ \leq \frac{2n}{m\delta} \mathbb{E}_{\infty,1} \left[\left| \frac{1}{n} \log \Lambda(1, n) - I \right| \right] \leq \frac{\epsilon}{2}, \end{aligned} \quad (4.85)$$

$$\begin{aligned} \mathbb{P}_{1,1} \left(\frac{n}{m} \left(\frac{1}{n} \log \Lambda(1, n) - I \right) \leq -\frac{\delta}{2} \right) \\ \leq \frac{2n}{m\delta} \mathbb{E}_{1,1} \left[\left| \frac{1}{n} \log \Lambda(1, n) - I \right| \right] \leq \frac{\epsilon}{2}. \end{aligned} \quad (4.86)$$

Next, we show that for any $t \geq T = 2N_1 + N_2$, both (4.81) and (4.82) are bounded by $\epsilon/2$. There are three possible cases:

1. $\nu_n - k \geq N_1$ and $t - (\nu_n - k) \geq N_1$,
2. $\nu_n - k < N_1$ and $t - (\nu_n - k) \geq N_1$,
3. $\nu_n - k \geq N_1$ and $t - (\nu_n - k) < N_1$.

Applying (4.83) and (4.84) in the first case, (4.84) and (4.85) in the second case, and (4.83) and (4.86) in the third case to (4.81) and (4.82), respectively, we obtain $\sup_{A_2} \mathbb{P}_{\nu_n, \nu_c} \left(\frac{1}{t} \log \Lambda(k, k + t - 1) - I \leq -\delta \right) \leq \epsilon$. The proof for (ii) is similar and proof is now complete.

Appendix 4.D Proof of Proposition 4.3

From (4.18), there exists $\gamma > 0$ such that $m_b \geq (1 + \gamma)b/I$ for all b sufficiently large. For any $0 < \epsilon < \gamma/(1 + \gamma)$, let $n_b = \left\lceil \frac{b}{(1-\epsilon)I} \right\rceil$ and $\delta = \epsilon I$. There exists $b_1 > 0$ such that $n_b(I - \delta) \geq b$ for all $b \geq b_1$. From Lemma 4.2, by choosing b_1 sufficiently large, we have for all $b \geq b_1$,

$$\begin{aligned} & \sup_{1 \leq \nu_c \leq k} \mathbb{P}_{\nu_n, \nu_c} (\log \Lambda(k, k + n_b - 1) < b) \\ & \leq \sup_{1 \leq \nu_c \leq k} \mathbb{P}_{\nu_n, \nu_c} (\log \Lambda(k, k + n_b - 1) \leq n_b(I - \delta)) \leq \epsilon. \end{aligned} \quad (4.87)$$

Let $b_2 \geq b_1$ be such that $I/b_2 \leq 1 + \gamma - (1 - \epsilon)^{-1}$. Then, for $b \geq b_2$, we have

$$\begin{aligned} \frac{m_b}{n_b} & \geq \frac{bI^{-1}(1 + \gamma)}{\frac{bI^{-1}}{1-\epsilon} + 1} \\ & \geq \frac{1 + \gamma}{(1 - \epsilon)^{-1} + I/b} \geq 1. \end{aligned}$$

For any $k, \nu_n, \nu_c \geq 1$, we then have

$$\begin{aligned} & \text{ess sup } \mathbb{P}_{\nu_n, \nu_c} (\tilde{\tau}_n(b) - \nu_c + 1 > kn_b \mid X_1, \dots, X_{\nu_c-1}) \\ & = \text{ess sup } \mathbb{P}_{\nu_n, \nu_c} \left(\max_{t-m_b \leq k' \leq t} \log \Lambda(k', t) < b \right. \\ & \quad \left. \text{for all } t \leq \nu_c + kn_b - 1 \mid X_1, \dots, X_{\nu_c-1} \right) \\ & \leq \text{ess sup } \mathbb{P}_{\nu_n, \nu_c} (\log \Lambda(\nu_c + (j-1)n_b, \nu_c + jn_b - 1) < b \\ & \quad \text{for all } 1 \leq j \leq k \mid X_1, \dots, X_{\nu_c-1}) \\ & = \prod_{j=1}^k \mathbb{P}_{\nu_n, \nu_c} (\log \Lambda(\nu_c + (j-1)n_b, \nu_c + jn_b - 1) < b) \leq \epsilon^k, \end{aligned}$$

where the last equality follows from independence and the last inequality from (4.87). Therefore, for any $b \geq b_2$, we have

$$\sup_{\nu_n, \nu_c \geq 1} \text{ess sup } \mathbb{E}_{\nu_n, \nu_c} [(\tilde{\tau}_n(b) - \nu_c + 1)^+ \mid X_1, \dots, X_{\nu_c-1}]$$

$$\begin{aligned}
&= \sup_{\nu_n, \nu_c \geq 1} \text{ess sup} \sum_{i=0}^{\infty} \mathbb{P}_{\nu_n, \nu_c} ((\tilde{\tau}_n(b) - \nu_c + 1) > i \mid X_1, \dots, X_{\nu_c-1}) \\
&\leq \sum_{k=0}^{\infty} n_b \sup_{\nu_n, \nu_c \geq 1} \text{ess sup} \mathbb{P}_{\nu_n, \nu_c} (\tilde{\tau}_n(b) - \nu_c + 1 > kn_b \mid X_1, \dots, X_{\nu_c-1}) \\
&\leq \frac{n_b}{1-\epsilon} \leq \frac{b}{I(1-\epsilon)^2} + \frac{1}{1-\epsilon} = b \left(I^{-1} + \frac{2\epsilon - \epsilon^2}{I(1-\epsilon)^2} + \frac{1}{b(1-\epsilon)} \right),
\end{aligned}$$

which yields (i). The proof for (ii) is similar and the proposition is proved.

Appendix 4.E Proof of Theorem 4.3

From Lemma 4.1, taking infimum on both sides of (4.31), we obtain $\text{ARL}(\tau_{\text{W-SGLR}}(b)) = \inf_{\nu_n \in \mathbb{N} \cup \{\infty\}} \mathbb{E}_{\nu_n, \infty} [\tau_{\text{W-SGLR}}(b)] \geq \frac{1}{2}e^b$. Since $\tau_{\text{W-SGLR}} \leq \tilde{\tau}$ and $\tau_{\text{W-SGLR}} \leq \tilde{\tau}_n$, by Proposition 4.3, we have $\text{WADD}(\tau_{\text{W-SGLR}}(b)) \leq (I^{-1} + o(1))b$ as $b \rightarrow \infty$.

To see that $\tau_{\text{W-SGLR}}(b)$ is asymptotically optimal when Assumption 4.2 is satisfied, let $C_\gamma = \{\tau : \text{ARL}(\tau) \geq \gamma\}$ be the set of stopping times satisfying $\text{ARL}(\tau) \geq \gamma$. By expanding $\text{WADD}(\tau)$ using (4.3), we obtain

$$\begin{aligned}
\inf_{\tau \in C_\gamma} \text{WADD}(\tau) &= \inf_{\tau \in C_\gamma} \sup_{\nu_n \in \mathbb{N} \cup \{\infty\}} \text{WADD}_{\nu_n}(\tau) \\
&\geq \sup_{\nu_n \in \mathbb{N} \cup \{\infty\}} \inf_{\tau \in C_\gamma} \text{WADD}_{\nu_n}(\tau) \tag{4.88}
\end{aligned}$$

$$\geq \sup_{\nu_n \in \{0, \nu_c, \infty\}} \inf_{\tau \in C_\gamma} \text{WADD}_{\nu_n}(\tau), \tag{4.89}$$

where (4.88) is due to the min-max inequality [126]. For each of the cases $\nu_n \in \{0, \nu_c, \infty\}$, by Theorem 6.17 in [18], we have

$$\begin{aligned}
\inf_{\tau \in C_\gamma} \text{WADD}_{\nu_n}(\tau) &\geq (D(g_n \| f_n)^{-1} + o(1)) b, \quad \text{when } \nu_n = 0, \\
\inf_{\tau \in C_\gamma} \text{WADD}_{\nu_n}(\tau) &\geq (D(g_n \| f)^{-1} + o(1)) b, \quad \text{when } \nu_n = \nu_c, \\
\inf_{\tau \in C_\gamma} \text{WADD}_{\nu_n}(\tau) &\geq (D(g \| f)^{-1} + o(1)) b, \quad \text{when } \nu_n = \infty.
\end{aligned}$$

Since Assumption 4.2 is satisfied, we have

$$I = \min \{D(g \parallel f), D(g_n \parallel f), D(g_n \parallel f_n)\}. \quad (4.90)$$

Therefore, from (4.89) and (4.90), we obtain

$$\begin{aligned} \inf_{\tau \in \mathcal{C}_\gamma} \text{WADD}_{\nu_n}(\tau) &\geq \left(\max \{D(g_n \parallel f_n)^{-1}, D(g_n \parallel f)^{-1}, \right. \\ &\quad \left. D(g \parallel f)^{-1}\} + o(1) \right) b \\ &= (I^{-1} + o(1)) b, \end{aligned}$$

and the proof is now complete.

Appendix 4.F Proof of Lemma 4.4

We use techniques is similar to [9] to prove Lemma 4.4. To analyze the probability $\mathbb{P}_{\nu_n, \infty}(\widehat{\eta}_k < \infty)$, we use a change-of-measure argument. For any $0 < \delta < 1$, choose $b_\delta \geq 0$ so that for any $b \geq b_\delta$

$$|\Theta| \pi^{-\frac{d}{2}} \Gamma\left(\frac{d}{2} + 1\right) b^{\frac{d}{2}} \exp(-b) \leq \exp(-(1 - \delta)b),$$

where $|\Theta|$ is the volume or Lebesgue measure of $\Theta \subset \mathbb{R}^d$ and $\Gamma(\cdot)$ is the gamma function. From Kolmogorov's Consistency Theorem, there is a probability measure G_θ for the stochastic process $(X_i)_{i \geq k}$ under which the pdf of each X_i is $g(\cdot; \theta)$. Define a measure $H(\cdot) = \int_\Theta G_\theta(\cdot) d\theta$. Since Θ is compact in $\mathbb{R}^d$, the measure H is finite. For each $t \geq k$, the Radon-Nikodym derivative of the law of $(X_k, X_{k+1}, \dots, X_t)$ under H w.r.t. $\mathbb{P}_{\nu_n, \infty}$ is

$$R_t = \int_\Theta \exp \left(\sum_{i=k}^t \log \frac{g(X_i; \theta)}{h_{\nu_n, \infty, \theta, \theta_n, i}(X_i)} \right) d\theta,$$

which follows from Fubini's Theorem. By Wald's likelihood ratio identity,

$$\begin{aligned}\mathbb{P}_{\nu_n\infty}(\widehat{\eta}_k < \infty) &= \int_{\{\widehat{\eta}_k < \infty\}} R_{\widehat{\eta}_k}^{-1} dH \\ &= \int_{\Theta} \left\{ \int_{\{\widehat{\eta}_k < \infty\}} R_{\widehat{\eta}_k}^{-1} dG_{\theta} \right\} d\theta.\end{aligned}\quad (4.91)$$

Suppose $\widehat{\eta}_k = t$. Since $\widehat{\theta} = \arg \max_{\theta} \log \widehat{\Lambda}(k, t, \theta) \in \text{Int}(\Theta)$, from Taylor series, there exists $\theta^* \in \Theta$ such that

$$\sum_{i=k}^t \log g(x_i; \theta) = \sum_{i=k}^t \log g(x_i; \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^T \left[\nabla^2 \sum_{i=k}^t \log g(x_i, \theta^*) \right] (\theta - \widehat{\theta}).$$

Thus, for $\|\theta - \widehat{\theta}\| < 1/\sqrt{b}$, we have

$$\begin{aligned}\log \widehat{\Lambda}(k, t, \widehat{\theta}) - \log \widehat{\Lambda}(k, t, \theta) &= -\frac{1}{2}(\theta - \widehat{\theta})^T \left[\nabla^2 \log \widehat{\Lambda}(k, t, \theta^*) \right] (\theta - \widehat{\theta}) \\ &\leq \frac{1}{2} \|\theta - \widehat{\theta}\|^2 \lambda_{\max} \left(-\nabla^2 \log \widehat{\Lambda}(k, t, \theta) \right) \\ &\leq \frac{1}{2} \|\theta - \widehat{\theta}\|^2 b,\end{aligned}$$

where the last inequality follows from $\sup_{\|\theta - \widehat{\theta}\| < 1/\sqrt{b}} \lambda_{\max} \left(-\nabla^2 \log \widehat{\Lambda}(k, t, \theta) \right) \leq b$.

We obtain

$$\begin{aligned}\frac{R_t}{\widehat{\Lambda}(k, t, \widehat{\theta})} &= \int_{\Theta} \exp \left(- \left[\log \widehat{\Lambda}(k, t, \widehat{\theta}) - \sum_{i=k}^t \log \frac{g(X_i, \theta)}{h_{\nu_n, \infty, \theta, \theta_n, i}(X_i)} \right] \right) d\theta \\ &\geq \int_{\Theta} \exp \left(- \left[\log \widehat{\Lambda}(k, t, \widehat{\theta}) - \log \widehat{\Lambda}(k, t, \theta) \right] \right) d\theta \\ &\geq \int_{\|\theta - \widehat{\theta}\| \leq 1/\sqrt{b}} \exp \left(-\frac{1}{2} \|\theta - \widehat{\theta}\|^2 b \right) d\theta \\ &\geq \int_{\|\theta - \widehat{\theta}\| \leq 1/\sqrt{b}} 1 d\theta \\ &\geq \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} b^{-\frac{d}{2}}\end{aligned}$$

Therefore, we have

$$\begin{aligned} R_t &\geq \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} b^{-\frac{d}{2}} \widehat{\Lambda}(k, t, \widehat{\theta}) \\ &\geq \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} b^{-\frac{d}{2}} \exp(b). \end{aligned}$$

This yields the upper bound

$$R_t^{-1} \leq \pi^{-\frac{d}{2}} \Gamma(\frac{d}{2} + 1) b^{\frac{d}{2}} \exp(-b).$$

Applying this upper bound to (4.91), we obtain

$$\begin{aligned} \mathbb{P}_{\nu_{n,\infty}}(\widehat{\eta}_k < \infty) &\leq \int_{\Theta} \left\{ \pi^{-\frac{d}{2}} \Gamma\left(\frac{d}{2} + 1\right) b^{\frac{d}{2}} \exp(-b) \int_{\{\widehat{\eta}_k < \infty\}} dG_{\theta} \right\} d\theta \\ &\leq \int_{\Theta} \left\{ \pi^{-\frac{d}{2}} \Gamma\left(\frac{d}{2} + 1\right) b^{\frac{d}{2}} \exp(-b) \right\} d\theta \\ &= |\Theta| \pi^{-\frac{d}{2}} \Gamma\left(\frac{d}{2} + 1\right) b^{\frac{d}{2}} \exp(-b) \\ &\leq \exp(-(1 - \delta)b), \end{aligned}$$

for all $b \geq b_{\delta}$. The proof that $\mathbb{P}_{\nu_{n,\infty}}(\widehat{\eta}_{n,l} < \infty) \leq \exp(-(1 - \delta)b)$ is similar, and the lemma is proved.

Chapter 5

QCD with sampling constraints

In this chapter, we discuss the QCD problem when sampling constraints are taken into consideration. In Section 5.1, we present our signal model and problem formulation. In Section 5.2, we present properties of the AOSC and the asymptotic ARL-WADD trade-off. In Section 5.3, we present the GLR CuSum stopping time and formulate the design of the sampling policy as a quadratic programming problem. We present algorithms to solve the policy design problems in Section 5.4. Numerical results are presented in Section 5.5.

Notations: For the rest of this chapter, we let $\mathbb{P}_{\nu,m}$ and $\mathbb{E}_{\nu,m}$ be the probability measure and mathematical expectation when the change point is at ν and post-change distribution is $\mathbb{P}^{f_m}$, respectively.

5.1 Problem formulation

Let $\mathbb{P}^{f_0}$ and $\mathbb{P}^{f_1}, \dots, \mathbb{P}^{f_M}$ be $M+1$ distinct distributions on $\mathbb{R}^N$, and $X_1, X_2, \dots$ be a sequence of vector-valued random variables satisfying the following:

$$\begin{cases} X_t \sim f_0 \text{ i.i.d. for all } t < \nu, \\ X_t \sim f_m \text{ i.i.d. for all } t \geq \nu. \end{cases} \quad (5.1)$$

where $\nu \geq 0$ and $m \in \{1, \dots, M\}$ are unknown but deterministic constants. The QCD problem is to detect the change in distribution as quickly as possible by observing $X_1, X_2, \dots$, while keeping the false alarm rate low.

In this chapter, we assume that the observer is only able to obtain a partial observation (A_t, Y_t) of X_t , where $Y_t \triangleq A_t(X_t)$ is a function of the random variable X_t under the action A_t at each time t . Let $\mathcal{A}$ be the collection of permissible actions. We assume that the set $\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$ is finite. We also assume that at each time t , under the distribution $\mathbb{P}^{f_m}$, the observation Y_t is conditionally independent of $Y_1, \dots, Y_{t-1}$ and $A_1, \dots, A_{t-1}$ given the action A_t . Some examples of $\mathcal{A}$ that arise in practical applications include:

1. Network Sampling. The set of rank $n < N$ transformations with

$$\mathcal{A} = \{ \mathbf{M} \mid \mathbf{M} = [e_{i_1}, \dots, e_{i_n}]^T, i_1 < i_2 < \dots < i_n \},$$

where e_i is an $N \times 1$ column vector with all zeros except a 1 at the i -th position, and $\mathbf{M}(X_t) = \mathbf{M}X_t$.

2. 1-bit Quantization. The set of functions on $\mathbb{R}$ for a fixed $n \in \mathbb{N}$ with

$$\mathcal{A} = \{ g : \mathbb{R} \rightarrow \mathbb{R} \mid g(x) = \mathbf{1}_{\{x \geq q\}}, q \in \{1, 2, \dots, 2^n\} \},$$

where $\mathbf{1}_A$ is the indicator function of the set A .

In our sequential change detection problem, we obtain observations $(A_1 = a_1, Y_1 = y_1), (A_2 = a_2, Y_2 = y_2), \dots$ sequentially and aim to detect the change in distribution from $\mathbb{P}^{f_0}$ to $\mathbb{P}^{f_m}$ for some fixed $m \in \{1, \dots, G\}$ as quickly as possible. This is determined by a stopping time. At each time instant, we also seek to find the best action to take based on actions from a past window of finite length.

Definition 5.1. A policy $\pi = (q, T)$ with window size $W < \infty$ is a Markov chain of order W on $\mathcal{A}$ with initial distribution q and probability transition matrix T .

For a stopping time τ and a policy π , we quantify its detection delay using the worst case average detection delay (WADD) as proposed by Lorden [7]:

$$\text{WADD}(\tau, \pi) = \max_{1 \leq m \leq M} \text{WADD}_m(\tau, \pi), \quad (5.2)$$

where

$$\begin{aligned} & \text{WADD}_m(\tau, \pi) \\ &= \sup_{\nu \geq 1} \text{ess sup}_{\mathbb{E}_{\nu, f_m}} [(\tau - \nu + 1)^+ \mid A^{1:(\nu-1)}, Y^{1:(\nu-1)}], \end{aligned}$$

and its ARL to false alarm as $\text{ARL}(\tau, \pi) = \mathbb{E}_{\infty}[\tau]$.

In order to take the observation-switching costs of a policy π into consideration, we let $\mathcal{C}$ be a $|\mathcal{A}| \times |\mathcal{A}|$ matrix where its (i, j) -th entry $\mathcal{C}[i, j]$ denotes the cost of switching from action i to action j . Inspired by a similar cost first proposed in [68], we define the AOSC of the policy π as

$$\text{AOSC}(\pi) = \limsup_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\infty} \left[\sum_{t=2}^{n+1} \mathcal{C}[A_{t-1}, A_t] \right]. \quad (5.3)$$

Formally, our quickest change detection with AOSC constraint can be formulated as a minimax problem: find a sampling policy π and a stopping time τ to

$$\begin{aligned} & \min_{\tau, \pi} \quad \text{WADD}(\tau, \pi) \\ & \text{s. t.} \quad \text{ARL}(\tau, \pi) \geq \gamma, \\ & \quad \text{AOSC}(\pi) \leq \alpha_{\text{AOSC}}, \end{aligned} \quad (5.4)$$

for some given thresholds α_{AOSC} and γ .

5.2 Properties of the AOSC

In this section, we present results that relate the AOSC and WADD of a sampling policy. When the window size of the sampling policy is zero, the actions used to observe the signal are generated independent and identically distributed (i.i.d.) with respect to the distribution q . In this case, the observations $\{(A_t, Y_t) = (a_t, y_t)\}_{n \in \mathbb{N}}$ are also generated i.i.d.. However, when the window size of the sampling policy is positive, unlike the former case, it is possible that the actions $\{A_t\}_{t \in \mathbb{N}}$ and observations $\{(A_t, Y_t)\}_{n \in \mathbb{N}}$ are not generated i.i.d.. We denote the joint probability distribution function of $(A^{1:t}, Y^{1:t})$ under f_m as

$$\begin{aligned} p_m(a^{1:t}, y^{1:t}) \\ = q(a^{1:M}) \prod_{j=1}^t p_m(y_j | a_j) \prod_{k=M+1}^t p_T(a_k | a^{k-M:k-1}), \end{aligned}$$

where p_T is the conditional probability mass function of A_k given $A^{k-M:k-1}$ induced by the probability transition matrix T and $p_m(y|a)$ is the conditional probability density function of an observation $Y_t = y$ given the action $A_t = a$ under the distribution $\mathbb{P}^{f_m}$.

A sampling policy $\pi = (q, T)$ with window size W can also be written as a Markov chain $\pi' = (q', T')$ of order 1 where T' satisfies $T'[\alpha, \beta] = 0$ whenever $\beta[i] \neq \alpha[i-1]$ for some $i \in \{2, \dots, W\}$ and $\alpha, \beta \in \mathcal{A}^W$. For the rest of this chapter, we switch between either representation of a sampling policy π to simplify the computations in the proofs. We denote the observation-switching costs associated with the latest two actions $\mathcal{C}[\alpha[W-1], \alpha[W]]$, from $\alpha[W-1]$ to $\alpha[W]$, as $\mathcal{C}_\alpha$.

For the rest of this section, we present results that relate the AOSC and WADD of sampling policies with different initial distributions but equal probability transition matrices. First, we recall a relation between the average number of visits and the stationary distributions of a Markov chain. Let $N_t(\alpha; \beta)$ denote the number of times, up to time t , that the state α is visited given that the initial state is β . Since $\mathcal{A}^W$ is finite, the Markov chain defined by the transition matrix T has at least one

recurrence class. Let R be the number of recurrence classes and U be the number of transient states. By the Ergodic Theorem for finite state Markov chains [127], for a finite state Markov Chain with R recurrent classes $\{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_R\}$, there exists R stationary distributions $\xi_1, \dots, \xi_R$ where $\xi_r[\alpha] = 0$ if the state $\alpha \notin \mathcal{R}_r$, and for recurrent states $\beta \in \mathcal{R}_r$, we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{N_t(\alpha; \beta)}{t} &= \xi_r[\alpha] \text{ a. s.}, \\ \lim_{t \rightarrow \infty} \frac{\mathbb{E}[N_t(\alpha; \beta)]}{t} &= \xi_r[\alpha], \end{aligned}$$

for any state $\alpha \in \mathcal{A}^W$. For transient states β ,

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}[N_t(\alpha; \beta)]}{t} = \sum_{r=1}^R f_{\beta,r} \xi_r[\alpha],$$

where $f_{\beta,r}$ is the first-passage probability of initializing at state β and entering the recurrence class $\mathcal{R}_r$ before any other recurrence classes.

For any state β , denoting ξ as the vector of expected proportion of visits to each of the states initializing at state β such that $\xi[\alpha] = \lim_{t \rightarrow \infty} \frac{\mathbb{E}[N_t(\alpha; \beta)]}{t}$, we can see that ξ is a stationary distribution of the probability transition matrix T as it is a convex linear combination of stationary distributions. Thus, for any initial distribution q , the expected proportion of visits to each of the states, denoted as $\bar{q}$, is a stationary distribution of T . In the next lemmas, we see that the AOSC and asymptotic log likelihood ratios depend only on T and the expected proportion of visits to each of the states.

Lemma 5.1. Let $\pi_1 = (q, T)$ be a policy of finite window size W and $\pi_2 = (\bar{q}, T)$, then we have

$$\text{AOSC}(\pi_1) = \text{AOSC}(\pi_2)$$

Proof. See Appendix 5.A. □

Next, we let

$$\Lambda_m(k, t) = \log \frac{p_m(A^{k:t}, Y^{k:t})}{p_0(A^{k:t}, Y^{k:t})} = \sum_{i=k}^t \log \frac{p_m(Y_i | A_i)}{p_0(Y_i | A_i)}.$$

We show that for any policy $\pi = (q, T)$ where $\bar{q}$ has support in only one recurrence class, the following results hold.

Lemma 5.2. For any policy $\pi = (q, T)$ of finite window size W where q has support in only one recurrence class $\mathcal{R}$, and any $m \in \{1, \dots, M\}$ and change-point $\nu < \infty$, we have

$$\mathbb{P}_{\nu, m} \left(\lim_{t \rightarrow \infty} \frac{1}{t} \Lambda_m(\nu, \nu + t) = I_{m, \pi} \right) = 1, \quad (5.5)$$

with

$$I_{m, \pi} = \sum_{\alpha \in \mathcal{R}} \bar{q}(\alpha) D(p_m(\cdot | \alpha[W]) \| p_0(\cdot | \alpha[W])). \quad (5.6)$$

Proof. See Appendix 5.B. □

Lemma 5.3. Let $\pi = (q, T)$ be a policy of finite window size W . Suppose $\bar{q}$ has support in only one recurrence class. For any $\epsilon > 0$ and $m \in \{1, \dots, M\}$, we have

$$\lim_{t \rightarrow \infty} \mathbb{P}_{\nu, m} \left(\left| \frac{1}{t} \Lambda_m(\nu, \nu + t) - I_{m, \pi} \right| > \epsilon \right) = 0, \quad (5.7)$$

for $0 \leq \nu < \infty$, and

$$\begin{aligned} \sup_{0 \leq \nu < \infty} \text{ess sup } \mathbb{P}_{\nu, m} \left(\frac{1}{t} \max_{0 \leq j < t} \Lambda_m(\nu, \nu + j) \right. \\ \left. > (1 + \epsilon) I_{m, \pi} \mid A^{1:\nu-1}, Y^{1:\nu-1} \right) \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Proof. See Appendix 5.C. □

5.3 Structure of Asymptotically Optimal Stopping time

In this section, we present the GLR CuSum test for the QCD problem with AOSC constraints and study its asymptotic properties as $\gamma \rightarrow \infty$.

5.3.1 Asymptotic Properties of GLR CuSum

For a fixed policy $\pi = (q, T)$, the GLR CuSum stopping time τ_π w.r.t. the observed sequence $(A_t = a_t, Y_t = y_t)_{t \in \mathbb{N}}$ is defined as:

$$\tau_\pi = \inf \{ t \mid S(t, \pi) > \log(M\gamma) \}, \quad (5.8)$$

$$S(t, \pi) = \max_{1 \leq m \leq M} \max_{1 \leq i \leq t+1} \sum_{j=i}^t \log \frac{p_m(y_j | a_j)}{p_0(y_j | a_j)}, \quad (5.9)$$

where $\gamma \in \mathbb{R}$ is a preselected threshold. The GLR CuSum stopping time τ_π can be re-written as

$$\begin{aligned} \tau_\pi &= \min_{1 \leq m \leq M} \tau_{m, \pi}, \quad S(t, \pi) = \max_{1 \leq m \leq M} S_m(t, \pi), \\ \tau_{m, \pi} &= \inf \{ t \mid S_m(t, \pi) > \log(M\gamma) \}, \\ S_m(t, \pi) &= \left(S_m(t-1, \pi) + \log \frac{p_m(y_t | a_t)}{p_0(y_t | a_t)}, 0 \right)^+ \text{ for } t > 0, \\ S_m(0, \pi) &= 0, \quad \text{for } m \in \{1, \dots, M\} \end{aligned}$$

where $x^+ \triangleq \max(x, 0)$. We note that for $m \in \{1, \dots, M\}$, $S_m(t, \pi)$ is the CuSum statistic corresponding to the post-change distribution $\mathbb{P}^{f_m}$ and policy π . Thus, the GLR CuSum statistic $S(t, \pi)$ is the maximum of the CuSum statistics for each of the post-change distributions $\mathbb{P}^{f_m}$. It should be noted that it is implicitly assumed in the formula of the GLR CuSum test statistic that the conditional distributions given the actions are known. This assumption is rather stringent and may be difficult to satisfy in many applications. One workaround for these applications is

to discretize the conditional distributions given the actions using the method of binning introduced in Chapter 3. Once the sample space is discretized, it is easy to use some training data to estimate the discretized conditional distributions given the actions and apply the GLR CuSum using the estimated distributions.

Next, we present some asymptotic properties of $S(t, \pi)$ and τ_π for a policy π . We use $\asymp$ to denote the notion of asymptotic equivalence [128]:

$$f \asymp g \quad \text{if and only if} \quad \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1.$$

For a fixed policy $\pi = (q, T)$ such that the expected proportion of visits to each of the states, $\bar{q}$, has support in one recurrence class, we apply Lemma 5.3 together with [16, Theorem 8.2.3] to obtain the following proposition.

Proposition 5.1. For a fixed policy $\pi = (q, T)$ such that $\bar{q}$ has support in one recurrence class, we have the following asymptotic ARL-WADD trade-off for any $1 \leq m \leq M$:

$$\text{ARL}(\tau_{m,\pi}, \pi) \geq M\gamma, \quad \text{WADD}_m(\tau_{m,\pi}, \pi) \leq \frac{\log \gamma}{I_{m,\pi}}(1 + o(1)),$$

as $\gamma \rightarrow \infty$.

Thus, when the signal is $\{X_t : t \in \mathbb{N}\}$ is sampled using the policy π such that $\bar{q}$ has support in one recurrence class, using similar techniques from [18, Theorem 6.16] together with Proposition 5.1, we know that the GLR CuSum stopping time τ_π gives us a stopping time satisfying $\text{ARL}(\tau_\pi, \pi) \geq \gamma$ and τ_π is asymptotically optimal for the following problem as $\gamma \rightarrow \infty$:

$$\min_{\tau} \text{WADD}(\tau, \pi) \quad \text{s.t.} \quad \text{ARL}(\tau, \pi) \geq \gamma. \quad (5.10)$$

In the next proposition, we discuss the case where $\bar{q}$ has support in multiple recurrence classes.

Proposition 5.2. Let the policy $\pi = (q, T)$ be such that $\bar{q}$ has support in multiple recurrence classes. Then, there exists a policy $\pi' = (q', T)$ where $\bar{q}'$ has support in only one recurrence class such that for any stopping time τ ,

$$\text{AOSC}(\pi') \leq \text{AOSC}(\pi) \text{ and } \text{WADD}(\tau, \pi') \leq \text{WADD}(\tau, \pi).$$

Proof. See Appendix 5.D. □

Using this proposition, we obtain a result regarding asymptotically optimal solutions of Problem (5.4).

Theorem 5.1. There exists a policy $\pi = (q', T')$ with $\bar{q}'$ having support in one recurrence class such that (τ_π, π) is asymptotically optimal for Problem (5.4) as $\gamma \rightarrow \infty$, where τ_π is the GLR CuSum stopping time.

Proof. See Appendix 5.E. □

Using Theorem 5.1, the minimization in Problem (5.4), over the sampling policy π and stopping time τ , can be decoupled. When the signal is sampled using a sampling policy $\pi = (q, T)$ with q having support in one recurrence class, satisfying $\text{AOSC}(\pi) \leq \alpha_{\text{AOSC}}$, the GLR CuSum τ_π is asymptotically optimal with the asymptotic WADD-ARL trade-off given as $\text{WADD}(\tau_\pi, \pi) = \frac{\log \gamma}{\min_m I_{m,\pi}} (1 + o(1))$ as $\gamma \rightarrow \infty$.

Definition 5.2. We call $\min_{1 \leq m \leq M} I_{m,\pi}$ the asymptotic ARL-WADD trade-off rate.

Let π^* be an optimal solution to the following problem:

$$\begin{aligned} \min_{\pi} \quad & \max_{1 \leq m \leq M} I_{m,\pi}^{-1} \\ \text{s. t.} \quad & \text{AOSC}(\pi) \leq \alpha_{\text{AOSC}}, \\ & \text{supp}(q) \subseteq \text{one recurrence class.} \end{aligned} \tag{5.11}$$

By the argument above, (τ_{π^*}, π^*) is asymptotically optimal for Problem (5.4). We call Problem (5.11) the *policy design* problem.

5.4 Optimal Sampling Policy

In this section, we investigate the sampling policy design Problem (5.11) under the cases where the switching costs from one action to another are all equal or not.

5.4.1 Equal Switching Costs

In this subsection, we propose a method to solve the policy design problem in which the switching costs are constant, i.e., $\mathcal{C}[a, b] = c$, for a fixed $c \in \mathbb{R}$ and any $a, b \in \mathcal{A}$. First, we note that Problem (5.11) is feasible if and only if $\alpha_{\text{AOSC}} \geq c$. Furthermore, if $\alpha_{\text{AOSC}} \geq c$ then Problem (5.4) reduces to

$$\begin{aligned} \min_{\tau, \pi} \quad & \text{WADD}(\tau, \pi) \\ \text{s. t.} \quad & \text{ARL}(\tau, \pi) \geq \gamma, \end{aligned} \tag{5.12}$$

where the AOSC constraint is automatically satisfied. Next, we show that for the case when all action-switching costs are equal, there exists a memoryless policy π (i.e., $W = 0$) for which the GLR CuSum τ_{π} achieves asymptotic optimality.

Proposition 5.3. Suppose Problem (5.11) is feasible and (τ_{π}, π) is an asymptotically optimal solution for Problem (5.4). There there exists a policy π_0 with window size $W = 0$ such that

$$\text{WADD}(\tau_{\pi_0}, \pi_0) \asymp \text{WADD}(\tau_{\pi}, \pi) \quad \text{as } \gamma \rightarrow \infty.$$

Proof. See Appendix 5.F. □

From Proposition 5.3, to solve the policy design problem Problem (5.11) for some $W \in \mathbb{N}$, it suffices to solve Problem (5.11) for the case where $W = 0$.

When $W = 0$, Problem (5.11) becomes

$$\begin{aligned} \min_q \quad & \max_{1 \leq m \leq M} \left(\sum_{a \in \mathcal{A}} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) \right)^{-1} \\ \text{s. t.} \quad & \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) \geq 0 \quad \text{for all } a \in \mathcal{A}, \end{aligned} \quad (5.13)$$

This is equivalent to solving the linear optimization problem:

$$\begin{aligned} \min_{q, x} \quad & x \\ \text{s. t.} \quad & \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\ & \sum_{a \in \mathcal{A}} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) + x \geq 0 \\ & \text{for all } m \in \{1, \dots, M\}. \end{aligned} \quad (5.14)$$

Let q_0 be the solution for Problem (5.14) and T_0 be the probability transition matrix with rows equal to q_0 . Using similar techniques from [18, Theorem 6.16] together with Proposition 5.1, we know that the GLR CuSum algorithm with $\pi_0 = (q_0, T_0)$ as the sampling policy gives us a stopping time satisfying $\text{ARL}(\tau_{q^*}) \geq \gamma$ with asymptotically optimal ARL-WADD trade-off as γ tends to infinity.

5.4.2 Unequal Switching Costs

In this subsection, we propose a method to solve the policy design problem in which the switching costs are not all equal. First, we present a proposition regarding the structure of asymptotically optimal solutions of Problem (5.4).

Proposition 5.4. Suppose (τ_π, π) is an asymptotically optimal solution for Problem (5.4) with window size at least 1. There exists a policy π_1 with window size $W = 1$ such that $\text{AOSC}(\pi_1) = \text{AOSC}(\pi)$ and

$$\text{WADD}(\tau_{\pi_1}, \pi_1) \asymp \text{WADD}(\tau_\pi, \pi) \quad \text{as } \gamma \rightarrow \infty.$$

Proof. See Appendix 5.G. □

From Proposition 5.4, the policy design problem for window size $W \in \mathbb{N}$ can be reduced to a problem of window size $\min(W, 1)$. Thus, we only need to study the cases where $W = 0$ or $W = 1$. In the following, we present algorithms to solve Problem (5.11) for each of these cases.

5.4.2.1 Window size $W = 0$

Using a similar argument from Section 5.4.1, we can see that for any optimal sampling policy $\pi = (q, T)$, we have $\bar{q} = q$ and T has only one recurrence class. Thus, we have

$$I_{m,\pi} = \sum_{a \in \mathcal{A}} q(a) D(p_m(\cdot | a) \| p_0(\cdot | a)),$$

$$\text{AOSC}(\pi) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} q(a) q(b) \mathcal{C}[a, b],$$

and Problem (5.11) becomes

$$\begin{aligned} \min_q \quad & \max_{1 \leq m \leq M} \left(\sum_{a \in \mathcal{A}} q(a) D(p_m(\cdot | a) \| p_0(\cdot | a)) \right)^{-1} \\ \text{s. t.} \quad & \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) \geq 0 \quad \text{for all } a \in \mathcal{A}, \\ & \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} q(a) q(b) \mathcal{C}[a, b] \leq \alpha_{\text{AOSC}}. \end{aligned} \tag{5.15}$$

Using the same argument from Problem (5.13), we can introduce a new variable x to obtain a linear cost function

$$\begin{aligned}
& \min_{q,x} \quad x \\
& \text{s. t.} \quad \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\
& \quad \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} q(a)q(b)\mathcal{C}[a, b] \leq \alpha_{\text{AOSC}}, \\
& \quad \sum_{a \in \mathcal{A}} q(a)D(p_0(\cdot | a) \| p_m(\cdot | a)) + x \geq 0 \\
& \quad \text{for all } m \in \{1, \dots, M\}.
\end{aligned} \tag{5.16}$$

This is a quadratically constrained quadratic program (QCQP), and we may assume that $\mathcal{C}$ is symmetric without loss of generality. However, without additional assumptions, the problem is NP-hard.

First, we discuss some special cases where the global optimum can be obtained easily. When $\mathcal{C}$ is positive semi-definite, Problem (5.16) is a convex programming problem. A convex program solver [129, 130] can be used to obtain globally optimal solutions. For the case where there are only cost of observations rather than cost of switching (i.e., $\mathcal{C}[a, b] = h(b)$ for some function $h : \mathcal{A} \rightarrow \mathbb{R}$), the quadratic constraint in Problem (5.16) reduces to

$$\sum_{b \in \mathcal{A}} q(b)h(b) \leq \alpha_{\text{AOSC}}. \tag{5.17}$$

In this case, Problem (5.16) becomes a linear programming problem, which can be solved by a linear program solver [129].

For the general case, we use the IRM algorithm [131] to obtain a locally optimal solution. In order to apply the IRM algorithm, we first rewrite the quadratic constraint

$$\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} q(a)q(b)\mathcal{C}[a, b] = q^T \mathcal{C} q = \text{Tr}(\mathcal{C} q q^T) \leq \alpha_{\text{AOSC}}$$

and Problem (5.16) is equivalent to

$$\begin{aligned}
& \min_{Q,q,x} \quad x \\
& \text{s. t.} \quad \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\
& \quad \text{Tr}(\mathcal{C}Q) \leq \alpha_{\text{AOSC}}, \\
& \quad Q\mathbf{1} = q, \\
& \quad Q \succeq 0 \quad \text{and} \quad \text{rank}(Q) = 1, \\
& \quad \sum_{a \in \mathcal{A}} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) + x \geq 0 \\
& \quad \text{for all } m \in \{1, \dots, M\},
\end{aligned} \tag{5.18}$$

where $\mathbf{1}$ is a $n^2 \times 1$ vector of ones. We note that Problem (5.18) becomes a convex programming problem when the constraint $\text{rank}(Q) = 1$ is ignored. We are now ready to present the IRM algorithm [131]. Fix $\omega > 1$.

First, we solve the convex problem

$$\begin{aligned}
& \min_{Q,q,x} \quad x \\
& \text{s. t.} \quad \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\
& \quad \text{Tr}(\mathcal{C}Q) \leq \alpha_{\text{AOSC}}, \\
& \quad Q\mathbf{1} = q, \\
& \quad Q \succeq 0, \\
& \quad \sum_{a \in \mathcal{A}} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) + x \geq 0 \\
& \quad \text{for all } m \in \{1, \dots, M\},
\end{aligned} \tag{5.19}$$

to obtain a solution (Q_0, q_0, x_0) and let $Q_0 = VDV^T$ be the eigen-decomposition of Q_0 . Let V_0 be the eigenvectors corresponding to the $n - 1$ smallest eigenvalues of Q_0 .

At each step k , we solve the following convex problem:

$$\begin{aligned}
& \min_{Q, q, x, r} && x + \omega^k r \\
& \text{s. t.} && \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\
& && \text{Tr}(\mathcal{C}Q) \leq \alpha_{\text{AOSC}}, \\
& && Q\mathbb{1} = q, \\
& && Q \succeq 0, \\
& && rI_{n-1} - V_{k-1}^T X V_{k-1} \succeq 0, \\
& && \sum_{a \in A} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) + x \geq 0 \\
& && \text{for all } m \in \{1, \dots, M\},
\end{aligned} \tag{5.20}$$

to obtain a solution (Q_k, q_k, x_k, r_k) and let V_k be the eigenvectors corresponding to the $n - 1$ smallest eigenvalues of Q_k . We iterate until $r_k < \epsilon$, where ϵ is a small threshold chosen as a stopping criterion. Following similar methods from [131], it can be shown that $r_k \rightarrow 0$ at a linear rate and that q_k converges to a locally optimal solution for Problem (5.18) if Problem (5.20) is feasible for all k .

5.4.2.2 Window size $W = 1$

Unlike the case when $W = 0$, not every distribution q is a stationary distribution of T . Furthermore, when $W > 0$, it is possible that more than one recurrence class

exists. In this case, Problem (5.11) becomes

$$\begin{aligned}
& \min_{T,q} \max_{1 \leq m \leq M} \left(\sum_{a \in \mathcal{A}} q(a) D(p_0(\cdot | a) \| p_m(\cdot | a)) \right)^{-1} \\
& \text{s. t. } \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0 \quad \text{for all } a \in \mathcal{A}, \\
& \sum_{b \in \mathcal{A}} T[a, b] = 1, \quad \text{for all } a \in \mathcal{A}, \\
& 0 \leq T[a, b] \leq 1, \quad \text{for all } a, b \in \mathcal{A}, \\
& q^T T = q^T, \\
& \sum_{b \in \mathcal{A}} \sum_{a \in \mathcal{A}} T[a, b] \mathcal{C}[a, b] q(a) \leq \alpha_{\text{AOSC}}, \\
& \text{supp}(q) \subseteq \text{one recurrence class.}
\end{aligned} \tag{5.21}$$

Using the same argument from Problem (5.13), Problem (5.21) is equivalent to

$$\begin{aligned}
& \min_{T,q,x} x \\
& \text{s. t. } \sum_{a \in \mathcal{A}} q(a) = 1, \quad q(a) > 0, \\
& \sum_{b \in \mathcal{A}} T[a, b] = 1, \\
& 0 \leq T[a, b] \leq 1, \\
& q^T T = q^T, \\
& \sum_{b \in \mathcal{A}} \sum_{a \in \mathcal{A}} T[a, b] \mathcal{C}[a, b] q(a) \leq \alpha_{\text{AOSC}}, \\
& \text{supp}(q) \subseteq \text{one recurrence class}, \\
& \sum_{a \in \mathcal{A}} q(a) D(p_m(\cdot | a) \| p_0(\cdot | a)) + x \geq 0, \\
& \text{for all } m \in \{1, \dots, M\}, \quad a, b \in \mathcal{A}.
\end{aligned} \tag{5.22}$$

Problem (5.22) has two quadratic constraints $Tq = q$ and $\sum_{b \in \mathcal{A}} \sum_{a \in \mathcal{A}} T[a, b] \mathcal{C}[a, b] q(b) \leq \alpha_{\text{AOSC}}$. Thus, it is a QCQP with an additional combinatorial constraint that $\text{supp}(q)$ is contained in a single recurrence class of T . Even without the combinatorial constraint, finding the global optimal for Problem (5.22) would be difficult

without additional assumptions on $\mathcal{C}$.

By considering a similar construction used in the proof of Proposition 5.4, we show that any policy $\pi_1 = (q, T)$ with window size 1 can be expressed as a policy $\pi_2 = (q_2, T_2)$ of window size 2 satisfying the following:

$$\sum_{a \in \mathcal{A}} q_2(a, b) = \sum_{c \in \mathcal{A}} q_2(c, b) \quad \text{for any } a, b \in \mathcal{A}, \quad (5.23)$$

$$T_2[(a, b), (c, d)] = \begin{cases} \frac{q_2(c, d)}{\sum_{d \in \mathcal{A}} q_2(c, d)} & \text{if } b = c, \\ 0 & \text{otherwise.} \end{cases} \quad (5.24)$$

Furthermore, any policy $\pi_2 = (q_2, T_2)$ with window size 2 that satisfies (5.23) and (5.24) can be expressed as a policy π_1 of window size 1. When we consider a policy $\pi_2 = (q_2, T_2)$ with window size 2 that satisfies (5.23) and (5.24), the constraint $T_2 q_2 = q_2$ becomes automatically satisfied and the AOSC constraint is linearized to $\sum_{(a,b) \in \mathcal{A}^2} \mathcal{C}[a, b] q(a, b) \leq \alpha_{\text{AOSC}}$. Hence, Problem (5.22) is equivalent to the following problem

$$\begin{aligned} \min_{q, x} \quad & x \\ \text{s. t.} \quad & \sum_{(a,b) \in \mathcal{A}^2} q(a, b) = 1, \quad q(a, b) > 0, \\ & \sum_{(a,b) \in \mathcal{A}^2} \mathcal{C}[a, b] q(a, b) \leq \alpha_{\text{AOSC}}, \\ & \sum_{a \in \mathcal{A}} q(a, b) = \sum_{c \in \mathcal{A}} q(c, b), \\ & \text{supp}(q) \subseteq \text{one recurrence class}, \\ & \sum_{(a,b) \in \mathcal{A}} q(a, b) D(p_m(\cdot | b) \| p_0(\cdot | b)) + x \geq 0, \\ & \text{for all } m \in \{1, \dots, M\}, \quad a, b \in \mathcal{A}, \end{aligned} \quad (5.25)$$

which is a linear programming problem with an additional combinatorial constraint. In order to handle the combinatorial constraint, we can solve Problem (5.25) without the combinatorial constraint $\text{supp}(q) \subseteq \text{one recurrence class}$ and if the solution

satisfies it, we have found the globally optimal solution. Alternatively, we can select a sufficiently small $\epsilon > 0$, and require that $q(a, b) > \epsilon$ for all $a, b \in \mathcal{A}$. The new constraint ensures that there is only one recurrence class for any feasible policy and thus, the recurrence class constraint is satisfied. The relaxed problem becomes

$$\begin{aligned}
& \min_{q, x} x \\
& \text{s. t.} \quad \sum_{(a, b) \in \mathcal{A}^2} q(a, b) = 1, \quad q(a, b) > \epsilon \text{ for all } a, b \in \mathcal{A}, \\
& \quad \sum_{(a, b) \in \mathcal{A}^2} \mathcal{C}[a, b] q(a, b) \leq \alpha_{\text{AOSC}}, \\
& \quad \sum_{a \in \mathcal{A}} q(a, b) = \sum_{c \in \mathcal{A}} q(c, b) \text{ for } b \in \mathcal{A}, \\
& \quad \sum_{(a, b) \in \mathcal{A}} q(a, b) D(p_m(\cdot | b) \parallel p_0(\cdot | b)) + x \geq 0, \\
& \quad \text{for all } m \in \{1, \dots, M\}, a, b \in \mathcal{A},
\end{aligned} \tag{5.26}$$

which can be solved by a linear program solver [129].

5.5 Numerical Results

In this section, we consider the QCD problem with AOSC based on partially observed graph signals under the various conditions discussed in this chapter. We consider the problem of quickest detection of a rogue node in a graph signal. We assume that our graph $\mathcal{G}$ is a connected graph with N nodes. We model the graph signal [132] in the pre-change regime with a zero-mean Gaussian distribution with covariance $\Sigma = L^\dagger + \epsilon I$, where $L^\dagger$ is the psuedo-inverse of the graph Laplacian matrix L , I is a $N \times N$ identity matrix and ϵ is the noise power in the graph signal. Thus, in the pre-change regime, we have

$$X_t \sim f_0 = \mathcal{N}(\mathbf{0}, \Sigma).$$

For the post-change regime, we assume that the signal obtained at the rogue node follows the same distribution as the pre-change distribution. However, this signal becomes independent of signals obtained from the rest of the graph. Thus, in the post-change regime, we have

$$X_t \sim f_m = \mathcal{N}(\mathbf{0}, \Sigma_m)$$

for $m \in \{1, \dots, M\}$, where the covariance matrix Σ_m is given as

$$\Sigma_m[i, j] = \begin{cases} \Sigma[i, j] & \text{for } i \neq m \text{ and } j \neq m, \\ \Sigma[m, m] & \text{for } i = j = m, \\ 0 & \text{otherwise.} \end{cases} \quad (5.27)$$

For the set of actions, we assume that the infrastructure constrains us to observe at most $N - 2$ nodes at any instance. The set of actions $\mathcal{A}$ is the set of all partial observations where

$$\mathcal{A} = \bigcup_{n=2}^{N-2} \{\mathbf{M} \mid \mathbf{M} = [e_{i_1}, e_{i_2}, \dots, e_{i_n}]^T, i_1 < i_2 < \dots < i_n\}.$$

In our experiments, we consider the graph $\mathcal{G}$, as shown in Fig. 5.1. In this case, there is a total of 50 actions that can be used to observe the signal on the graph $\mathcal{G}$.

Topology of Graph $\mathcal{G}$

FIG. 5.1: Graph $\mathcal{G}$ generated using the Erdős-Rényi (ER) random graph model with $N = 6$ nodes and probability of an edge $p = 0.2$.

We consider three possible cases of observation-switching costs $\mathcal{C}$:

1. no costs are involved with $\mathcal{C}^1 = \mathbf{0}$;
2. there is only activation cost where $\mathcal{C}^2[i, j]$ is the number of nodes observed using action i ;
3. there are both switching and activation costs with $\mathcal{C}^3 = \mathcal{C}^2 + \mathcal{C}'$ where $\mathcal{C}'[i, j]$ is the number of elements in the symmetric difference between the set of nodes observed by action i and set of nodes observed by action j .

Note that each $\mathcal{C}^1, \mathcal{C}^2, \mathcal{C}^3 \in \mathbb{R}^{50 \times 50}$.

When $\mathcal{C} = \mathcal{C}^1$, using results from Section 5.4.1, we only need to design a policy of window size 0. By considering the problem for $W = 0$, we obtain the optimal policy $\pi_{1, W=0}$ by solving Problem (5.14). Similarly, for the cases where $\mathcal{C} = \mathcal{C}^2, \mathcal{C}^3$, using results from Section 5.4.2, we only need to consider the cases $W = 0, 1$. When

$\mathcal{C} = \mathcal{C}^2$ and $W = 0$, we obtain the optimal policy $\pi_{2,W=0}$ by solving Problem (5.16) with a linear AOSC constraint (5.17). For the case $\mathcal{C} = \mathcal{C}^3$ with $W = 0$, we obtain a locally optimal policy $\pi_{3,W=0}$ by performing the IRM algorithm described in Section 5.4.2. In the case where $\mathcal{C} = \mathcal{C}^2$ or $\mathcal{C}^3$ with $W = 1$, we obtain an approximately optimal policy $\pi_{2,W=1,\epsilon}$ or $\pi_{3,W=1,\epsilon}$ respectively by solving Problem (5.26) with $\epsilon \in \{10^{-4}, 10^{-5}, 10^{-6}\}$.

When no costs are involved (i.e., $\mathcal{C} = \mathbf{0}$), the optimal asymptotic ARL-WADD trade-off rate (cf. Definition 5.2) achieved by the stopping rule and policy $(\tau_{\pi_{1,W=0}}, \pi_{1,W=0})$ is $\bar{I} = 1.5586$. This is an upper bound for the asymptotic ARL-WADD trade-off rates for all other choices of $\mathcal{C}$.

In Fig. 5.2, we compare the asymptotic ARL-WADD trade-off rate of $\pi_{2,W=0}$ and $\pi_{2,W=1,\epsilon}$. First, we observe that the asymptotic ARL-WADD trade-off is below the upper bound $\bar{I}$ across the range of achievable AOSC. In the case where $\mathcal{C} = \mathcal{C}^2$, the AOSC constraint in Problem (5.16) reduces to

$$\text{AOSC}(\pi) = \sum_{b \in \mathcal{A}} \left(\sum_{a \in \mathcal{A}} q(a, b) \right) h(b)$$

where $h(b)$ is the number of nodes observed using action b . Using similar arguments from the proof of Proposition 5.3, we can see that there is an asymptotically optimal trade-off for a policy window of size $W = 0$ is equal to the asymptotically optimal trade-off for a policy of window size $W = 1$. Hence, the performance between $\pi_{2,W=0}$ and $\pi_{2,W=1,\epsilon}$ becomes more similar as ϵ tends to zero.

In Fig. 5.3, we compare the asymptotic ARL-WADD trade-off rate of $\pi_{3,W=0}$ and $\pi_{3,W=1,\epsilon}$. Similarly, we observe that the asymptotic ARL-WADD trade-off is below the upper bound $\bar{I}$ across the range of achievable AOSC. In this case, we observe that the optimal policies of window size $W = 1$ significantly outperforms the optimal policies of window size $W = 0$. When we are using a policy of window size $W = 0$, the lowest AOSC achievable while Problem (5.4) remains feasible is about 4. However, using a policy of window size $W = 1$, we are able to reduce AOSC to

FIG. 5.2: Comparison of the ARL-WADD trade-off rate for different stopping times when $C = C_2$.

about 2 while Problem (5.4) remains feasible. These can be seen by comparing the lowest AOSC achieved by the respective curves in Fig. 5.3.

FIG. 5.3: Comparison of the ARL-WADD trade-off rate for different stopping times when $C = C_3$.

Appendix 5.A Proof of Lemma 5.1

Each time a state α is visited, a switching cost $\mathcal{C}_\alpha$ is incurred. Thus, the AOSC can be written as

$$\begin{aligned} \text{AOSC}(\pi_1) &= \lim_{t \rightarrow \infty} \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} \mathbb{E}_\infty \left[\frac{\mathcal{C}_\alpha N_t(\alpha; \beta)}{t} q(\beta) \right] \\ &= \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} q(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right]. \end{aligned} \quad (5.28)$$

Similarly, for π_2 , we have

$$\text{AOSC}(\pi_2) = \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} \bar{q}(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right]. \quad (5.29)$$

Suppose there are R recurrence classes $\mathcal{R}_1, \dots, \mathcal{R}_R$. For any $r \in \{1, \dots, R\}$ with both $\beta, \delta \in \mathcal{R}_r$, we have

$$\lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] = \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \delta)}{t} \right]. \quad (5.30)$$

Let $\beta_r \in \mathcal{R}_r$ for $r = 1, \dots, R$ and for any transient state δ , we have

$$\lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \delta)}{t} \right] = \sum_{r=1}^R f_{\delta, r} \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta_r)}{t} \right]. \quad (5.31)$$

Let $\{\delta_j \mid j = 1, \dots, U\}$ be the transient states. For any $r = 1, \dots, R$, we have

$$\sum_{\beta \in \mathcal{R}_r} \bar{q}(\beta) = \sum_{\beta \in \mathcal{R}_r} q(\beta) + \sum_{j=1}^U f_{\delta_j, r} q(\delta_j). \quad (5.32)$$

Combining equations (5.28) to (5.32) together, yields

$$\begin{aligned} \text{AOSC}(\pi_2) &= \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} \bar{q}(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \end{aligned} \quad (5.33)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{A}^W} \bar{q}(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \quad (5.34)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{R}_r} \bar{q}(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \quad (5.35)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \bar{q}(\beta) \sum_{\alpha \in \mathcal{R}_r} \mathcal{C}_\alpha \xi_r[\alpha] \quad (5.36)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} q(\beta) \sum_{\alpha \in \mathcal{R}_r} \mathcal{C}_\alpha \xi_r[\alpha] \\ + \sum_{r=1}^R \sum_{j=1}^U f_{\delta_j, r} q(\delta_j) \sum_{\alpha \in \mathcal{R}_r} \mathcal{C}_\alpha \xi_r[\alpha] \quad (5.37)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{R}_r} q(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \\ + \sum_{r=1}^R \sum_{j=1}^U \sum_{\alpha \in \mathcal{R}_r} f_{\delta_j, r} q(\delta_j) \mathcal{C}_\alpha \xi_r[\alpha] \quad (5.38)$$

where (5.34) is due to $\xi_r[\alpha] = 0$ for $\alpha \notin \mathcal{R}_r$, (5.37) is obtained by applying (5.32) to (5.36) and (5.38) is obtained by applying (5.30) to (5.37). We also have

$$\text{AOSC}(\pi_1) \\ = \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} q(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \quad (5.39)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{A}^W} q(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \\ + \sum_{j=1}^U \sum_{\alpha \in \mathcal{A}^W} q(\delta_j) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \delta_j)}{t} \right] \quad (5.40)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{A}^W} q(\beta) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right] \\ + \sum_{r=1}^R \sum_{j=1}^U \sum_{\alpha \in \mathcal{A}^W} f_{\delta_j, r} q(\delta_j) \mathcal{C}_\alpha \xi_r[\alpha] \quad (5.41)$$

$$= \sum_{r=1}^R \sum_{\beta \in \mathcal{R}_r} \sum_{\alpha \in \mathcal{A}^W} q(\delta_j) \mathcal{C}_\alpha \lim_{t \rightarrow \infty} \mathbb{E}_\infty \left[\frac{N_t(\alpha; \beta)}{t} \right]$$

$$+ \sum_{r=1}^R \sum_{j=1}^U \sum_{\alpha \in \mathcal{R}_r} f_{\delta_j, r} q(\delta_j) \mathcal{C}_\alpha \xi_r[\alpha], \quad (5.42)$$

where (5.40) is obtained by partitioning the actions in $\mathcal{A}^W$ into recurrence classes $\mathcal{R}_1, \dots, \mathcal{R}_R$ and transient states, (5.41) is obtained by applying (5.31) to (5.40) and, (5.42) is due to $\xi_r[\alpha] = 0$ for $\alpha \notin \mathcal{R}_r$. Thus, we have $\text{AOSC}(\pi_1) = \text{AOSC}(\pi_2)$ and the proof is complete.

Appendix 5.B Proof of Lemma 5.2

Since $\text{supp}(q)$ lies in the recurrence class $\mathcal{R}$, for any $t \geq \nu$ and $\beta \in \mathcal{A}^W$ such that $q(\beta) > 0$, we have $\beta \in \mathcal{R}$ and

$$\begin{aligned} & \frac{1}{t - \nu + 1} \Lambda_m(\nu, t) \\ &= \sum_{\alpha \in \mathcal{A}^W} \frac{1}{t - \nu + 1} \sum_{\left\{ \substack{i \text{ s.t. } \nu \leq i \leq t, \\ A^{i-W+1:i} = \alpha} \right\}} \log \frac{p_m(Y_i | \alpha[W])}{p_0(Y_i | \alpha[W])} \\ &= \sum_{\alpha \in \mathcal{R}} \frac{N_t(\alpha; \beta)}{t - \nu + 1} \frac{1}{N_t(\alpha; \beta)} \sum_{\left\{ \substack{i \text{ s.t. } \nu \leq i \leq t, \\ A^{i-W+1:i} = \alpha} \right\}} \log \frac{p_m(Y_t | \alpha[W])}{p_0(Y_t | \alpha[W])}, \end{aligned} \quad (5.43)$$

where the last equality follows because for any $\alpha \notin \mathcal{R}$, $\left\{ \substack{i \text{ s.t. } \nu \leq i \leq t, \\ A^{i-W+1:i} = \alpha} \right\} = \emptyset$. We have

$$\frac{N_t(\alpha; \beta)}{t - \nu + 1} \xrightarrow[t \rightarrow \infty]{\mathbb{P}_{\nu, m} \text{ a.s.}} \bar{q}(\alpha) \quad \text{as } t \rightarrow \infty,$$

and

$$\begin{aligned} & \frac{1}{N_t(\alpha; \beta)} \sum_{\left\{ \substack{i \text{ s.t. } \nu \leq i \leq t, \\ A^{i-W+1:i} = \alpha} \right\}} \log \frac{p_m(Y_t | \alpha[W])}{p_0(Y_t | \alpha[W])} \\ & \xrightarrow[t \rightarrow \infty]{\mathbb{P}_{\nu, m} \text{ a.s.}} D(p_m(\cdot | \alpha[W]) \| p_0(\cdot | \alpha[W])) \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Thus from (5.43), we obtain

$$\begin{aligned} & \frac{1}{t - \nu + 1} \Lambda_m(k, t) \\ & \xrightarrow[\text{a.s.}]{\mathbb{P}_{\nu, m}} \sum_{\alpha \in \mathcal{R}} \bar{q}(\alpha) D(p_m(\cdot \mid \alpha[W]) \parallel p_0(\cdot \mid \alpha[W])) = I_{m, \pi}. \end{aligned}$$

The proof is now complete.

Appendix 5.C Proof of Lemma 5.3

Suppose $\text{supp}(\bar{q}) \subseteq \mathcal{R}$ for a recurrence class $\mathcal{R}$, then $\text{supp}(q) \subseteq \mathcal{R} \cup \mathcal{U}$, where $\mathcal{U}$ is a set of transient states such that the first-passage probability of entering $\mathcal{R}$ from each $\beta \in \mathcal{U}$ is one. Then, from Lemma 5.2, (5.7) follows. Next, we have

$$\begin{aligned} & \sup_{0 \leq \nu < \infty} \text{ess sup } \mathbb{P}_{\nu, m} \left(t^{-1} \max_{0 \leq j < t} \Lambda_m(\nu, \nu + j) \right. \\ & \quad \left. > (1 + \epsilon) I_{m, \pi} \mid A^{1:\nu-1}, Y^{1:\nu-1} \right) \end{aligned} \quad (5.44)$$

$$\begin{aligned} & = \sup_{0 \leq \nu < \infty} \text{ess sup } \mathbb{P}_{\nu, m} \left(t^{-1} \max_{0 \leq j < t} \Lambda_m(\nu, \nu + j) \right. \\ & \quad \left. > (1 + \epsilon) I_{m, \pi} \mid A^{\nu-W:\nu-1} \right) \end{aligned} \quad (5.45)$$

$$\begin{aligned} & = \sup_{0 \leq \nu < \infty} \max_{\substack{\alpha \in \mathcal{A}^W \\ q(\alpha) > 0}} \mathbb{P}_{\nu, m} \left(t^{-1} \max_{0 \leq j < t} \Lambda_m(\nu, \nu + j) \right. \\ & \quad \left. > (1 + \epsilon) I_{m, \pi} \mid A^{\nu-W:\nu-1} = \alpha \right) \\ & = \max_{\substack{\alpha \in \mathcal{A}^W \\ q(\alpha) > 0}} \mathbb{P}_{W+1, m} \left(t^{-1} \max_{0 \leq j < t} \Lambda_m(W + 1, W + 1 + j) \right. \\ & \quad \left. > (1 + \epsilon) I_{m, \pi} \mid A^{1:W} = \alpha \right) \rightarrow 0 \quad \text{as } t \rightarrow \infty, \end{aligned} \quad (5.46)$$

where (5.45) is because $\Lambda_m(\nu, \nu + j)$ is independent of $A^{1:\nu-W-1}$ and $Y^{1:\nu-1}$ given $A^{\nu-W:\nu-1}$ for each $0 \leq j < t$, and (5.46) is because each of the terms within the set $\{\alpha \in \mathcal{A}^W \mid q(\alpha) > 0\}$ that we take maximum over converges to zero due to Lemma 5.2.

Appendix 5.D Proof of Proposition 5.2

Without loss of generality, we consider the case where $\bar{q}$ has support in the recurrence classes $\mathcal{R}_1$ and $\mathcal{R}_2$. Let q_1, q_2 be stationary distributions on $\mathcal{A}^W$ such that q_i has support only in $\mathcal{R}_i$ for $i \in \{1, 2\}$ and that there exists $0 < \lambda < 1$ such that $\bar{q} = \lambda q_1 + (1 - \lambda)q_2$. Let $\pi_1 = (q_1, T)$ and $\pi_2 = (q_2, T)$. From Lemma 5.1, we obtain

$$\begin{aligned} \text{AOSC}(q, T) &= \text{AOSC}(\bar{q}, T) \\ &= \lambda \text{AOSC}(q_1, T) + (1 - \lambda) \text{AOSC}(q_2, T) \\ &\geq \min \{ \text{AOSC}(\pi_1), \text{AOSC}(\pi_2) \}. \end{aligned}$$

Let E_1 be the event that $\mathcal{R}_1$ is visited before $\mathcal{R}_2$ and E_2 be the event that $\mathcal{R}_2$ is visited before $\mathcal{R}_1$. We have

$$\begin{aligned} \text{WADD}(\tau, \pi) &= \sup_{\substack{\nu \geq 1 \\ 1 \leq m \leq M}} \text{ess sup } \mathbb{E}_{\nu, f_m} [(\tau - \nu + 1)^+ \mid A^{1:\nu-1}, Y^{1:\nu-1}] \\ &= \sup_{\substack{\nu \geq 1 \\ 1 \leq m \leq M}} \max_{i=1,2} \text{ess sup } \mathbb{E}_{\nu, f_m} [(\tau - \nu + 1)^+ \mid A^{1:\nu-1}, Y^{1:\nu-1}, E_i] \\ &= \max_{i=1,2} \sup_{\substack{\nu \geq 1 \\ 1 \leq m \leq M}} \text{ess sup } \mathbb{E}_{\nu, f_m} [(\tau - \nu + 1)^+ \mid A^{1:\nu-1}, Y^{1:\nu-1}, E_i] \\ &= \max_{i=1,2} \sup_{\substack{\nu \geq 1 \\ 1 \leq m \leq M}} \text{ess sup } \mathbb{E}_{\nu, f_m} [(\tau - \nu + 1)^+ \mid A^{1:\nu-1}, Y^{1:\nu-1}, E_i] \\ &= \max_{i=1,2} \text{WADD}(\tau, \pi_i). \end{aligned}$$

Setting $\pi' = \pi_j$ where $j = \arg \min_{i=1,2} \text{AOSC}(\pi_i)$ completes the proof.

Appendix 5.E Proof of Theorem 5.1

We first note that Problem (5.4) is feasible and an optimal solution exists. Let the policy $\pi^* = (q, T)$ and stopping-time τ^* such that (τ^*, π^*) is optimal for Problem (5.4). From Proposition 5.2, we can find a policy $\pi = (q', T)$ such that $\bar{q}'$ has support in only one recurrence class and

$$\begin{aligned} \text{AOSC}(\pi) &\leq \text{AOSC}(\pi^*), \\ \text{WADD}(\tau^*, \pi) &\leq \text{WADD}(\tau^*, \pi^*). \end{aligned}$$

From the discussion before (5.10), the GLR CuSum stopping time τ_π is asymptotically optimal for the following problem:

$$\min_{\tau} \text{WADD}(\tau, \pi) \quad \text{s.t.} \quad \text{ARL}(\tau, \pi) \geq \gamma. \quad (5.47)$$

Thus, the GLR CuSum stopping time τ_π satisfies

$$\text{WADD}(\tau_\pi, \pi) \asymp \text{WADD}(\tau^*, \pi) \leq \text{WADD}(\tau^*, \pi^*),$$

as $\gamma \rightarrow \infty$. Hence, (τ_π, π) is also an asymptotically optimal solution to Problem (5.4).

Appendix 5.F Proof of Proposition 5.3

If π is equivalent to a policy of window size 0, then $\pi_0 = \pi$ proves the proposition. If $\pi = (q, T)$ is a policy of window size $W > 0$, by Theorem 5.1, it suffices to consider the case where the support of $\bar{q}$ is a subset of a recurrence class of T . Let $\pi_0 = (q_0, T_0)$ such that $q_0 = \bar{q}^0$ and T_0 be the probability transition matrix with rows equal to q_0 where

$$\bar{q}^0(a) = \sum_{\alpha \in \mathcal{A}^W: \alpha[W]=a} \bar{q}(\alpha)$$

for any $a \in \mathcal{A}$. For any $m \in \{1, \dots, M\}$, we have

$$\begin{aligned}
I_{m,\pi} &= \sum_{\alpha \in \mathcal{A}^W} \bar{q}(\alpha) D(p_m(\cdot \mid \alpha[W]) \parallel p_0(\cdot \mid \alpha[W])) \\
&= \sum_{a \in \mathcal{A}} \sum_{\substack{\alpha \in \mathcal{A}^W: \\ \alpha[W]=a}} \bar{q}(\alpha) D(p_m(\cdot \mid \alpha[W]) \parallel p_0(\cdot \mid \alpha[W])) \\
&= \sum_{a \in \mathcal{A}} D(p_m(\cdot \mid a) \parallel p_0(\cdot \mid a)) \sum_{\substack{\alpha \in \mathcal{A}^W: \\ \alpha[W]=a}} \bar{q}(\alpha) \\
&= \sum_{a \in \mathcal{A}} q_0(a) D(p_m(\cdot \mid a) \parallel p_0(\cdot \mid a)) = I_{m,\pi_0}, \tag{5.48}
\end{aligned}$$

where (5.48) is due to the fact that $\bar{q}_0 = q_0$ as π_0 is a policy of window size 0. By Proposition 5.1, we obtain $\text{WADD}(\tau_\pi, \pi) \asymp \text{WADD}(\tau_{\pi_0}, \pi_0)$ as $\gamma \rightarrow \infty$. The proof is now complete.

Appendix 5.G Proof of Proposition 5.4

If π is equivalent to a policy of window size 1, we let $\pi_1 = \pi$ and we have proved the proposition. If π is a policy of window size $W > 1$, by Theorem 5.1, it suffices to consider the case where the support of $\bar{q}$ is a subset of a recurrence class $\mathcal{R}$. The AOSC of π can be expressed as

$$\begin{aligned}
&\text{AOSC}(\pi) \\
&= \text{AOSC}(\bar{q}, T) \\
&= \limsup_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_\infty \left[\sum_{t=2}^{n+1} \mathcal{C}[A_{i-1}, A_i] \right] \\
&= \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{\beta \in \mathcal{A}^W} \bar{q}(\beta) \mathbb{E}_\infty \left[\sum_{t=2}^{n+1} \mathcal{C}[A_{i-1}, A_i] \mid A^{1:W} = \beta \right] \\
&= \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{\beta \in \mathcal{A}^W} \bar{q}(\beta) \mathbb{E}_\infty \left[\sum_{\alpha \in \mathcal{A}^W} \mathcal{C}_\alpha N_{n+1}(\alpha; \beta) \mid A^{1:W} = \beta \right] \\
&= \sum_{\beta \in \mathcal{A}^W} \sum_{\alpha \in \mathcal{A}^W} \bar{q}(\beta) \mathcal{C}_\alpha \bar{q}(\alpha) = \sum_{\alpha \in \mathcal{A}^W} \mathcal{C}_\alpha \bar{q}(\alpha).
\end{aligned}$$

Let $\pi_2 = (q_2, T_2)$ be a policy of window size 2 and it can be similarly shown that $\text{AOSC}(\pi_2) = \sum_{\alpha \in \mathcal{A}^2} \mathcal{C}_\alpha \bar{q}_2(\alpha)$. Thus, it can be seen that the $\text{AOSC}(\pi_2)$ depends only on $\bar{q}_2$. Let $q_2 = \bar{q}^1$ be the projection of $\bar{q}$ onto $\mathcal{A}^2$ where

$$\bar{q}^1(a, b) = \sum_{\substack{\alpha \in \mathcal{A}^W: \\ \alpha[W-1]=a, \alpha[W]=b}} \bar{q}(\alpha),$$

for any $a, b \in \mathcal{A}$ and T_2 be defined as

$$T_2[(a, b), (c, d)] = \begin{cases} \frac{\bar{q}^1(c, d)}{\sum_{d \in \mathcal{A}} \bar{q}^1(c, d)} & \text{if } c = b, \\ 0 & \text{otherwise.} \end{cases}$$

For any $(a, b) \in \mathcal{A}^2$, we have

$$\begin{aligned} \sum_{(c, d) \in \mathcal{A}^2} T_2[(a, b), (c, d)] &= \sum_{d \in \mathcal{A}} \frac{\bar{q}^1(b, d)}{\sum_{d_1 \in \mathcal{A}} \bar{q}^1(b, d_1)} \\ &= \frac{\sum_{d \in \mathcal{A}} \bar{q}^1(b, d)}{\sum_{d_1 \in \mathcal{A}} \bar{q}^1(b, d_1)} = 1, \end{aligned}$$

and hence, T_2 is a probability transition matrix. Since the support of $\bar{q}$ is a subset of a single recurrence class of T , the support of q_2 also lies in a single recurrence class of T_2 . Next, we show that q_2 is a stationary distribution of T_2 :

$$\begin{aligned} T_2 q_2[(c, d)] &= \sum_{(a, b) \in \mathcal{A}^2} T_2[(a, b), (c, d)] \bar{q}^1(a, b) \\ &= \sum_{a \in \mathcal{A}} T_2[(a, c), (c, d)] \bar{q}^1(a, c) \\ &= \sum_{a \in \mathcal{A}} \frac{\bar{q}^1(c, d)}{\sum_{d \in \mathcal{A}} \bar{q}^1(c, d)} \bar{q}^1(a, c) \\ &= \frac{\bar{q}^1(c, d)}{\sum_{d \in \mathcal{A}} \bar{q}^1(c, d)} \sum_{a \in \mathcal{A}} \bar{q}^1(a, c) \\ &= \bar{q}^1(c, d) = q_2(c, d). \end{aligned}$$

Therefore, q_2 is a stationary distribution of T_2 which has support contained in one recurrence class. Hence, $\overline{q_2} = q_2$ and $\text{AOSC}(\pi) = \text{AOSC}(\pi_2)$. Using similar arguments from the proof of Proposition 5.3, we have $I_{m,\pi} = I_{m,\pi_1}$ for $m = \{1, \dots, M\}$ and $\text{WADD}(\tau_\pi, \pi) \asymp \text{WADD}(\tau_{\pi_2}, \pi_2)$ as $\gamma \rightarrow \infty$.

Furthermore, a quick computation shows that $\pi_2 = (q_2, T_2)$ is equivalent to the policy $\pi_1 = (q_1, T_1)$ of window size 1 with $q_1(a) = \sum_{b \in \mathcal{A}} q_2(a, b)$ and $T_1[a, b] = \frac{q_2(a, b)}{\sum_{b \in \mathcal{A}} q_2(a, b)}$.

Thus, we have $\text{AOSC}(\pi) = \text{AOSC}(\pi_1)$ and $\text{WADD}(\tau_\pi, \pi) \asymp \text{WADD}(\tau_{\pi_1}, \pi_1)$ as $\gamma \rightarrow \infty$. The proof is complete.

Chapter 6

Privacy-Aware Quickest Change Detection

In this chapter, we discuss the QCD problem while taking privacy constraints into considerations. In Section 6.1, we present our signal model and problem formulation. We derive the asymptotic optimality of the GLR CuSum stopping time and formulate an optimization problem to design the optimal sanitization channel in Section 6.2. We propose relaxations to the channel design problem in Section 6.3 and present methods to solve the relaxed channel design problem in Section 6.4. Results from numerical experiments are presented in Section 6.5.

Notations: For the rest of this chapter, we denote $\mathbb{E}_{\nu,i}$ as the expectation operator assuming the change-point is at ν with post-change distribution is $\mathbb{P}^{g_i}$ and $\mathbb{E}_{\infty}$ as the expectation operator assuming the change does not occur.

6.1 Problem formulation

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, where $\mathcal{X}$ is a finite alphabet. Let $\mathbb{P}^f$ be the pre-change distribution and $G = \{\mathbb{P}^{g_1}, \mathbb{P}^{g_2}, \dots, \mathbb{P}^{g_{|G|}}\}$ be the set of possible post-change distributions such that $\mathbb{P}^f \neq \mathbb{P}^{g_i}$ for all $i \in \{1, 2, \dots, |G|\}$. Let I be a random

variable on the indices of G with distribution p_I . Let $X_1, X_2, \dots$, be a sequence of random variables taking values in $\mathcal{X}$, and satisfying the following:

$$\begin{cases} X_t \sim f & \text{i.i.d. for all } t < \nu, \\ X_t \sim g_i & \text{i.i.d. for all } t \geq \nu, \end{cases} \quad (6.1)$$

where $\nu \geq 0$ is an unknown but deterministic change point, i is the realization of the random variable I , which remains the same for all $t \geq \nu$. We further assume that there is an unknown random variable U , which takes on values in a finite set $\mathcal{U}$, and that U can be expressed as a randomized function of I , i.e., which distribution is the post-change distribution tells us information about U which an adversary might be interested in.

A sanitization channel is a sequence of function $\{q_t\}_{t \in \mathbb{N}}$ which the function q_t maps the observations up to time t , $X_1, X_2, \dots, X_t$ to a random variable Y_t . Examples of sanitization channels include non-stationary sanitization channels $Y_t = q_t(x_t)$, sanitization channels with full memory $Y_t = q_t(x_1, \dots, x_t)$ and windowed sanitization channels $Y_t = q(x_{t-W+1}, \dots, x_t)$ where W is the window size. To simplify our analysis of the privacy-aware QCD problem, we restrict our analysis to a memoryless privacy mechanism and leave more general privacy mechanisms for a future study. A memoryless privacy mapping or sanitization channel q maps an observation $X \in \mathcal{X}$ to a random variable $Y \in \mathcal{Y}$, where $\mathcal{Y}$ is a discrete alphabet. The privacy mechanism q can be represented by the conditional probability $\mathbb{P}(Y = y | X = x)$. Let T_q be a column-stochastic matrix with $[T_q]_{y,x} = \mathbb{P}(Y = y | X = x)$ where $[T]_{y,x}$ denotes the (y, x) entry of a matrix T . Likewise, we represent a distribution h on $\mathcal{X}$ as a column vector with $[h]_x = h(x)$ and similarly for a distribution on $\mathcal{Y}$.

At each time t , we apply a sanitization channel q to obtain $Y_t = q(X_t)$. The sanitized signal Y_t is generated i.i.d. by the distribution $\mathbb{P}^{T_q f}$ in the pre-change regime and by the distribution $\mathbb{P}^{T_q g_i}$ in the post-change regime, for some $1 \leq i \leq |G|$. For a fixed q , we let $\tilde{G} = \{\mathbb{P}^{\tilde{g}} : \mathbb{P}^{\tilde{g}} = \mathbb{P}^{T_q g_i} \text{ for } 1 \leq i \leq |G|\}$ to be the set of

possible post-change distributions and $\mathbb{P}^{\tilde{f}} = \mathbb{P}^{T_q f}$ to be the pre-change distribution of the sanitized signal, respectively.

In this chapter, we study the QCD problem with privacy constraints using two different privacy metrics. We assume that, at each time t , the adversary knows the pre-change distribution $\mathbb{P}^f$, the set G of post-change distributions, change-point ν , the sanitization channel q , the current and all previous observations $y^{1:t} = \{y_1, y_2, \dots, y_t\}$.

The first privacy metric we consider is the concept of maximal leakage, first proposed in [91], to quantify the amount of information leakage an adversary is able to gain from observing the signal $\{Y_t : t \in \mathbb{N}\}$. Given two random variables A and B , the maximal leakage from A to B is defined as

$$\mathcal{L}_{\max}(A \rightarrow B) = \sup_{U-A-B-\hat{U}} \log \frac{\mathbb{P}(\hat{U} = U)}{\max_{u \in \mathcal{U}} \mathbb{P}(U = u)}, \quad (6.2)$$

where $U-A-B-\hat{U}$ denotes a Markov chain and the supremum is taken over all such Markov chains. The quantity $\mathcal{L}_{\max}(A \rightarrow B)$ can be interpreted as the maximum gain in bits an adversary can achieve in guessing U by observing B , where U is a randomized function of A . The expression in (6.2) is equivalent [91] to

$$\mathcal{L}_{\max}(A \rightarrow B) = \log_2 \left(\sum_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \mathbb{P}(b | a) \right). \quad (6.3)$$

Thus, at each time t , the maximum gain in bits an adversary can achieve in guessing U is $\mathcal{L}_{\max}(I \rightarrow Y^{\nu:t})$.

The second privacy metric we consider is the *sequential hypothesis testing privacy* metric. It quantifies the amount of gain an adversary is able to achieve by performing a sequential hypothesis test. Using this privacy metric, we are interested in protecting a subset of the post-change hypotheses from the inference of a potential

adversary while ensuring the distinguishability of the rest of the post-change hypotheses. Given a sanitization channel q and a partition $G_1 \cup G_2$ of G with $|G_1| > 1$, such that G_1 is the set of post-change hypotheses which we wish to protect from the inference of a potential adversary, we define the *privacy of G_1 in G* and the *distinguishability of G_2 in G* as $\mathcal{K}_1(T_q)$ and $\mathcal{K}_2(T_q)$, respectively, such that

$$\begin{aligned}\mathcal{K}_1(T_q) &= \max_{\mathbb{P}^g \in G_1} \min_{\mathbb{P}^h \in G_1} D(T_q g \parallel T_q h), \\ \mathcal{K}_2(T_q) &= \min_{\mathbb{P}^g \in G_2} \min_{\mathbb{P}^h \in G} D(T_q g \parallel T_q h).\end{aligned}$$

Using standard results from sequential hypothesis testing [16], assuming that the adversary is willing to accept a misclassification rate of η , the expected number of samples required to identify a distribution in G_1 is at least $|\log \eta|/\mathcal{K}_1(T_q)$ asymptotically as $\eta \rightarrow 0$. Similarly, the expected number of samples required to identify a distribution in G_2 is $|\log \eta|/\mathcal{K}_2(T_q)$ asymptotically as $\eta \rightarrow 0$.

For a fixed q , the QCD problem is to detect a change in distribution as quickly as possible by observing the sanitized signal $Y_1 = q(X_1), Y_2 = q(X_2), \dots$, while keeping the false alarm rate low. In a typical sequential change detection procedure, at each time t , a test statistic $S(t)$ is computed based on the currently available observations $Y_1 = y_1, \dots, Y_t = y_t$, and the observer decides that a change has occurred at a stopping time $\tau = \inf\{t : S(t) > b\}$, which is the first t such that $S(t)$ exceeds a pre-determined threshold b . The QCD performance of a stopping time τ can be quantified using the trade-off between two quantities, the average run length to false alarm, $\text{ARL}(\tau)$, and the expected worst-case average detection delay, $\text{EWADD}(\tau)$, with

$$\begin{aligned}\text{ARL}(\tau) &= \mathbb{E}_\infty [\tau], \quad \text{EWADD}(\tau) = \mathbb{E} [\text{WADD}_I(\tau)], \\ \text{WADD}_i(\tau) &= \sup_{\nu \geq 1} \text{ess sup } \mathbb{E}_{\nu, i} [(\tau - \nu + 1)^+ | Y_1^{\nu-1}],\end{aligned}$$

where ess sup is the essential supremum operator. We should note that, typically, there are several sanitization channels q which satisfies the privacy constraint. As the pre and post-change distributions, $\{\mathbb{P}^{\tilde{f}}\} \cup \tilde{G}$, vary with the sanitization channel

q , we expect that the QCD performance varies with q as well. It is then important for us to select the sanitization channel q that provides the best QCD performance while satisfying the privacy constraint.

Our privacy-aware QCD problem can be formulated as an optimization problem as follows: given a privacy admissible set $\mathcal{Q}$ and an average run length requirement γ , we seek a stopping time τ , and a sanitization channel q such that they are optimal solutions to the following problem:

$$\begin{aligned} & \underset{\tau, q}{\text{minimize}} && \text{EWADD}(\tau) \\ & \text{s. t.} && \text{ARL}(\tau) \geq \gamma, \\ & && q \in \mathcal{Q} \end{aligned} \tag{6.4}$$

where the privacy admissible set $\mathcal{Q} = \{q : \sup_{0 \leq \nu \leq t} \mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \epsilon\}$ for some given $\epsilon > 0$ under the maximal leakage privacy metric, and $\mathcal{Q} = \{q : \mathcal{K}_1(T_q) \leq \epsilon_1, \mathcal{K}_2(T_q) \geq \epsilon_2\}$ for some given $\epsilon_1, \epsilon_2 > 0$ under the sequential hypothesis testing privacy metric.

6.2 Asymptotic optimality

In this section, we present the GLR CuSum stopping time for the privacy-aware QCD problem and study its asymptotic properties as $\gamma \rightarrow \infty$. First, we note that the minimization over the sanitization channel q and stopping time τ can be decoupled in Problem (6.4). For a fixed $q \in \mathcal{Q}$, we define the GLR CuSum stopping time ω_q and the GLR CuSum test statistic $S(t)$ as follows

$$\begin{aligned} \omega_q &= \inf \{t : S(t) \geq b\}, \quad S(t) = \max_{1 \leq j \leq |\tilde{G}|} S_j(t), \\ \begin{cases} S_j(t) &= \max \left(S_j(t-1) + \log \frac{\tilde{g}_j(y_t)}{\tilde{f}(y_t)}, 0 \right) \\ S_j(0) &= 0, \end{cases} \quad \text{for } 1 \leq j \leq |\tilde{G}|. \end{aligned}$$

When the signal $\{x_t : t \in \mathbb{N}\}$ is sanitized using the channel q , the GLR CuSum stopping time ω_q is asymptotically optimal [7] for the following problem:

$$\underset{\tau}{\text{minimize}} \quad \text{EWADD}(\tau) \quad \text{s. t.} \quad \text{ARL}(\tau) \geq \gamma, \quad (6.5)$$

with the asymptotic ARL-EWADD trade-off given as $\text{EWADD}(\omega_q) = \mathbb{E} \left[\frac{\log \gamma}{D(T_q g_I \| T_q f)} \right] (1 + o(1))$ as $\gamma \rightarrow \infty$, where the expectation is taken with respect to I . Let q^* be an optimal solution to the following problem:

$$\begin{aligned} &\underset{q}{\text{minimize}} \quad \mathbb{E} \left[\frac{1}{D(T_q g_I \| T_q f)} \right] \\ &\text{s. t.} \quad q \in \mathcal{Q}, \end{aligned} \quad (6.6)$$

Using similar arguments from [14, 16], ω_{q^*} is asymptotically optimal for Problem (6.4) as $\gamma \rightarrow \infty$.

We call Problem (6.6) the channel design problem and note that it is challenging to solve for several reasons. First, the objective function is neither concave nor convex. Second, under the maximal leakage privacy metric, it is difficult to use (6.3) to compute a closed form expression for the privacy constraint since the alphabet size of $Y^{\nu:t}$ increases quickly as $t \rightarrow \infty$. Under the sequential hypothesis testing privacy metric, the constraints $\mathcal{K}_1(T_q) \leq \epsilon_1$ and $\mathcal{K}_2(T_q) \geq \epsilon_2$ are neither concave nor convex. In the next section, we present relaxations of the constraint and objective function of Problem (6.6) to improve its computational tractability.

6.3 Relaxation of the Channel Design Problem

6.3.1 Relaxation of the Objective Function

In this subsection, we provide a relaxation of the objective function in Problem (6.6). We propose to relax the objective function using Jensen's inequality $\mathbb{E} \left[\frac{1}{D(T_q g_I \| T_q f)} \right] \geq \frac{1}{\mathbb{E}[D(T_q g_I \| T_q f)]}$. By replacing the objective function with its lower

bound, we obtain the following relaxed problem:

$$\begin{aligned} & \underset{q}{\text{maximize}} && \mathbb{E}[D(T_q g_I \| T_q f)] \\ & \text{s. t.} && q \in \mathcal{Q}. \end{aligned} \tag{6.7}$$

For a fixed post-change distribution $\mathbb{P}^{g_i}$, with $1 \leq i \leq N$, the expected rate of growth of $S(t)$ in the post-change regime is given as $D(T_q g_i \| T_q f)$. Thus, we can interpret the new objective function as the expected rate of growth of $S(t)$ averaged over the different post-change distributions. Since the stopping time ω_q declares that a change has taken place when the test statistic $S(t)$ exceeds a pre-defined threshold b , this means that, heuristically, a larger expected rate of growth of $S(t)$ gives a smaller EWADD. This intuition agrees with the relaxed problem (6.7) obtained by replacing the objective function with its lower bound $1/\mathbb{E}[D(T_q g_I \| T_q f)]$ in Problem (6.6).

6.3.2 Relaxation of Privacy Constraints

6.3.2.1 Maximal Leakage privacy

In this subsection, we focus on the relaxation of the constraint $q \in \mathcal{Q}$ when the privacy metric is the maximal leakage privacy. Under the maximal leakage privacy metric, Problem (6.7) becomes:

$$\begin{aligned} & \underset{q}{\text{maximize}} && \mathbb{E}[D(T_q g_I \| T_q f)] \\ & \text{s. t.} && \sup_{0 \leq \nu \leq t} \mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \epsilon. \end{aligned} \tag{6.8}$$

As it is difficult to obtain a closed form expression for the maximal leakage $\mathcal{L}(I \rightarrow Y^{\nu:t})$, we approximate it using an upper bound which is easily computable. Let J

be a random variable on the indices of $\tilde{G} = \{\tilde{g}_1, \tilde{g}_2, \dots, \tilde{g}_{|\tilde{G}|}\}$, such that

$$\mathbb{P}(J = j | I = i) = \begin{cases} 1 & \text{if } \tilde{g}_j = T_q g_i, \\ 0 & \text{otherwise.} \end{cases}$$

Let U be a randomized function of I , according to our signal model, we have the following factorization,

$$\begin{aligned} P_{J,I,Y^{\nu:t},U} &= P_{J,I,U} P_{Y^{\nu:t}|I,J,U} \\ &= P_{I,J,U} P_{Y^{\nu:t}|J} \\ &= P_I P_{J|I} P_{U|I} P_{Y^{\nu:t}|J} \end{aligned}$$

where we use P_X to denote the probability mass function of X and $P_{X|Y}$ to denote the conditional probability mass function of X given Y . The following proposition provides motivation to relax the privacy constraint in Problem (6.6) to $\mathcal{L}_{\max}(I \rightarrow J) \leq \epsilon$.

Proposition 6.1. For any $t \in \mathbb{N}$, we have $\mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \mathcal{L}_{\max}(I \rightarrow J)$. Hence, we have

$$\sup_{0 \leq \nu \leq t} \mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \mathcal{L}_{\max}(I \rightarrow J).$$

Proof. We define the following matrices $\Delta \in \mathbb{R}^{|G| \times |\mathcal{U}|}$, $\Theta \in \mathbb{R}^{|\mathcal{Y}|^{t-\nu+1} \times |\tilde{G}|}$, $\Lambda \in \mathbb{R}^{|\tilde{G}| \times |G|}$, $\Phi_0 \in \mathbb{R}^{|U| \times |\mathcal{Y}|^{t-\nu+1}}$ and $\Phi_1 \in \mathbb{R}^{|U| \times |\tilde{G}|}$ such that

$$[\Delta]_{i,u} = \mathbb{P}(u | i) \mathbb{P}(i), \quad [\Theta]_{y^{\nu:t}, j} = \mathbb{P}(y^{\nu:t} | j), \quad (6.9)$$

$$[\Lambda]_{j,i} = \mathbb{P}(j | i), \quad (6.10)$$

and

$$[\Phi_0]_{u, y^{\nu:t}} = \begin{cases} 1 & \text{if } u = \arg \max_{u \in \mathcal{U}} \mathbb{P}(u, y^{\nu:t}), \\ 0 & \text{otherwise.} \end{cases} \quad (6.11)$$

$$[\Phi_1]_{u,j} = \begin{cases} 1 & \text{if } u = \arg \max_{u \in \mathcal{U}} \mathbb{P}(u, j), \\ 0 & \text{otherwise.} \end{cases} \quad (6.12)$$

Let U be a randomized function of I . We have

$$\sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \mathbb{P}(u, y^{\nu:t}) \quad (6.13)$$

$$= \sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \sum_{j=1}^{|\tilde{G}|} \sum_{i=1}^{|G|} \mathbb{P}(j, i, y^{\nu:t}, u) \quad (6.14)$$

$$= \sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \sum_{i=1}^{|G|} \mathbb{P}(u | i) \mathbb{P}(i) \left(\sum_{j=1}^{|\tilde{G}|} \mathbb{P}(y^{\nu:t} | j) \mathbb{P}(j | i) \right). \quad (6.15)$$

Similarly, we have

$$\sum_{j=1}^{|\tilde{G}|} \max_{u \in \mathcal{U}} \mathbb{P}(u, j) = \sum_{j=1}^{|\tilde{G}|} \max_{u \in \mathcal{U}} \sum_{y^{\nu:t}} \sum_{i=1}^{|G|} \mathbb{P}(j, i, y^{\nu:t}, u) \quad (6.16)$$

$$= \sum_{j=1}^{|\tilde{G}|} \max_{u \in \mathcal{U}} \sum_{i=1}^{|G|} \mathbb{P}(u | i) \mathbb{P}(i) \mathbb{P}(j | i). \quad (6.17)$$

From (6.9) and (6.11), we can see that (6.13) and (6.16) can be expressed using matrix operations

$$\begin{aligned} \sum_{j=1}^{|\tilde{G}|} \max_{u \in \mathcal{U}} \mathbb{P}(u, j) &= \text{Trace}(\Lambda \Delta \Phi_1), \\ \sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \mathbb{P}(u, y^{\nu:t}) &= \text{Trace}(\Theta \Lambda \Delta \Phi_0), \end{aligned}$$

Furthermore, Φ_1 is a solution to the following problem, $\Phi_1 = \arg \max_{\Gamma} \text{Trace}(\Lambda \Delta \Gamma)$, where the maximization is taken over all column stochastic matrices Γ . Since Φ_0, Θ are column stochastic matrices, we have $\text{Trace}(\Lambda \Delta \Phi_0 \Theta) \leq \text{Trace}(\Lambda \Delta \Phi_1)$, and thus $\sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \mathbb{P}(u, y^{\nu:t}) \leq \sum_{j=1}^{|\tilde{G}|} \max_{u \in \mathcal{U}} \mathbb{P}(u, j)$. Using the following equations [91],

$$\mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) = \sup_U \frac{\sum_{y^{\nu:t}} \max_{u \in \mathcal{U}} \mathbb{P}(u, y^{\nu:t})}{\max_{u \in \mathcal{U}} \mathbb{P}(u)}$$

$$\mathcal{L}_{\max}(I \rightarrow J) = \sup_U \frac{\sum_{\phi} \max_{u \in \mathcal{U}} \mathbb{P}(u, m)}{\max_{u \in \mathcal{U}} \mathbb{P}(u)}.$$

to compute maximal leakage, where the supremum is taken over all randomized functions U of I , we have $\mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \mathcal{L}_{\max}(I \rightarrow J)$. The proof is complete. $\square$

Replacing the privacy constraint in (6.7) with $\mathcal{L}_{\max}(I \rightarrow J)$ to obtain the relaxed channel design problem:

$$\begin{aligned} & \underset{q}{\text{maximize}} && \mathbb{E}[D(T_q g_I \| T_q f)] \\ & \text{s. t.} && \mathcal{L}_{\max}(I \rightarrow J) \leq \epsilon. \end{aligned} \tag{6.18}$$

Proposition 6.1 shows that, any solution to Problem (6.18) satisfies the original privacy constraint $\sup_{0 \leq \nu \leq t} \mathcal{L}_{\max}(I \rightarrow Y^{\nu:t}) \leq \epsilon$.

6.3.2.2 Relaxation of the sequential hypothesis testing privacy constraint

In this subsection, we focus on the relaxation of the constraint $q \in \mathcal{Q}$ when the privacy metric is the sequential hypothesis testing privacy metric. Under the sequential hypothesis testing privacy metric, Problem (6.7) becomes:

$$\begin{aligned} & \underset{q}{\text{maximize}} && \mathbb{E}[D(T_q g_I \| T_q f)] \\ & \text{s. t.} && \mathcal{K}_1(T_q) \leq \epsilon_1, \mathcal{K}_2(T_q) \geq \epsilon_2. \end{aligned} \tag{6.19}$$

Problem (6.19) is non-convex as the constraints $\mathcal{K}_1(T_q) \leq \epsilon_1$ and $\mathcal{K}_2(T_q) \geq \epsilon_2$ are non-convex. In this case, we focus on a restricted sanitization model in order to improve the computational tractability of Problem (6.19). Let $\mathcal{C} = \{q_1, q_2, \dots, q_n\}$ be a finite set of sanitization channels with column-stochastic matrices $T_1, \dots, T_n$. At each time instance t , we assume that the observer obtains an observation $Y_t = (Z_t, A_t)$ where $Z_t = q_{A_t}(X_t)$ is a randomized function of the random variable X_t under the sanitization channel q_{A_t} . We further assume that $\{A_t\}_{t \in \mathbb{N}}$ are

i.i.d. generated with distribution $\vec{q}$. Under the restricted sanitization model, rather than designing the sanitization channel q , we design the distribution $\vec{q}$ that samples a sanitization channel q_{A_t} from $\mathcal{C}$ at each time instance t and use q_{A_t} to sanitize the signal.

Using similar arguments from [14], the asymptotic ARL-EWADD trade-off of the GLR CuSum stopping time under the restricted sanitization model is given as EWADD(ω_q)

$= \mathbb{E} \left[\frac{\log \gamma}{\sum_{i=1}^n \vec{q}(i) D(T_i g_I \| T_i f)} \right] (1 + o(1))$ as $\gamma \rightarrow \infty$. Thus, Problem (6.19) becomes

$$\begin{aligned} & \underset{\vec{q}}{\text{maximize}} && \sum_{i=1}^n \vec{q}(i) \mathbb{E} [D(T_i g_I \| T_i f)] \\ & \text{s. t.} && \max_{g \in G_1} \min_{h \in G_1} \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \leq \epsilon_1, \\ & && \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \geq \epsilon_2 \text{ for } g \in G_2 \text{ and } h \in G. \end{aligned} \tag{6.20}$$

Since $D(T_i g_I \| T_i f)$, $D(T_i g \| T_i h)$ can be pre-computed for all $i \in \{1, \dots, n\}$ and $g, h \in G$, Problem (6.20) without the constraint $\max_{g \in G_1} \min_{h \in G_1} \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \leq \epsilon_1$ is a linear program. In the next section, we will show that Problem (6.20) can be expressed as a mixed-integer linear program (MILP).

6.4 Algorithms

In this section, exact and locally optimal methods for Problems (6.18) and (6.20) are presented.

6.4.1 Maximal Leakage Privacy

6.4.1.1 Exact Method

Since the conditional probability $\mathbb{P}(J = j | I = i) \in \{0, 1\}$, from (6.3), $2^{\mathcal{L}_{\max}(I \rightarrow J)}$ is an integer and the constraint in Problem (6.18) is equivalent to

$$\sum_j \max_i \mathbb{P}(J = j | I = i) \leq m, \quad (6.21)$$

where $m = \lfloor 2^\epsilon \rfloor$. Together with the fact that $P_{J|I}$ is a conditional probability represented by a zero-one matrix, we can deduce that for each $i = 1, \dots, |G|$, there exist $j_1, j_2, \dots, j_m$, not necessarily distinct, such that $j \in \{j_1, j_2, \dots, j_m\}$ whenever $\mathbb{P}(J = j | I = i) = 1$. Therefore, there is a total of $m^{|G|}$ conditional distributions that satisfy (6.21). For each conditional distribution $P_{J|I}$, Problem (6.18) is equivalent to the following problem:

$$\begin{aligned} & \underset{q}{\text{maximize}} \quad \mathbb{E}[D(T_q g_i \| T_q f)] \\ & \text{s. t.} \quad T_q g_i = \tilde{g}_j \\ & \quad \text{for all } i, j \text{ such that } \mathbb{P}(J = j | I = i) = 1 \end{aligned} \quad (6.22)$$

For a fixed conditional probability $P_{J|I}$, Problem (6.22) is maximizing a convex function over a convex bounded polytope. Therefore, an extreme point achieves the maximum value, and we are able to solve Problem (6.22) by enumerating over the finite number of extreme points on the convex bounded polytope. A solution for Problem (6.18) can be obtained by enumerating over all conditional probabilities $P_{J|I}$ represented by a zero-one matrix satisfying (6.21), and solving Problem (6.22) for each of these conditional probabilities.

By considering the symmetry of the labels of the sanitized post-change distributions, rather than $m^{|G|}$ different cases, we would only need to solve Problem (6.22) for $\left\{ \begin{smallmatrix} |G| \\ m \end{smallmatrix} \right\} < m^{|G|}$ different cases, where the Stirling number of the second kind [133] $\left\{ \begin{smallmatrix} a \\ b \end{smallmatrix} \right\}$ counts the number of ways to partition a set of a labeled objects into b

nonempty unlabeled subsets. However, we still need to solve at least exponentially many convex maximization problems with respect to the number of post-change distributions $|G|$, since $\left\{\frac{|G|}{m}\right\} \sim \frac{m^{|G|}}{m!}$ as $|G| \rightarrow \infty$. This maybe computationally undesirable when $|G|$ is large.

6.4.1.2 Augmented Lagrangian Method

In this subsection, we further relax the channel design problem by relaxing the discrete constraint (6.21). This relaxation allows the application of the augmented Lagrangian method for cases where the exact method is computationally undesirable. The constraint (6.21) is equivalent to $|\{T_q g_i : 1 \leq i \leq |G|\}| \leq m$. In order to count the number of distinct elements in the set $\{T_q g_i : 1 \leq i \leq |G|\}$, we can use the following relaxation,

$$\begin{aligned} |\{T_q g_i : 1 \leq i \leq |G|\}| &= 1 + \sum_{i=2}^{|G|} \prod_{j=1}^{i-1} \mathbf{1}_{T_q g_i \neq T_q g_j} \\ &\approx 1 + \sum_{i=2}^{|G|} \prod_{j=1}^{i-1} \left(\frac{1}{2} + \frac{1}{\pi} \arctan(k \|T_q g_i - T_q g_j\|_1) \right), \end{aligned}$$

where k is a chosen to be large and $\|\cdot\|_1$ refers to the vector 1-norm. Putting this back into Problem (6.18), we obtain the following continuous optimization problem:

$$\begin{aligned} &\underset{q}{\text{maximize}} \quad \mathbb{E}[D(T_q g_n \| T_q f)] \\ &\text{s. t.} \quad \sum_{i=2}^{|G|} \prod_{j=1}^{i-1} \left(\frac{1}{2} + \frac{1}{\pi} \arctan(k \|T_q g_i - T_q g_j\|_1) \right) \leq m - 1, \end{aligned} \tag{6.23}$$

which an augmented Lagrangian Solver [134, 135] can be used to obtain locally optimal [136] solutions.

6.4.2 Sequential Hypothesis Testing Privacy

In this subsection, we show that Problem (6.20) is equivalent to a MILP. First, we require the following proposition.

Proposition 6.2. For distribution $\vec{q}$ on $\{1, \dots, n\}$, $\vec{q}$ satisfies

$$\max_{g \in G_1} \min_{h \in G_1} \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \leq \epsilon_1 \quad (6.24)$$

if and only if there exist functions $\xi : G_1 \rightarrow \mathbb{R}$ and $\delta : G_1 \times G_1 \rightarrow \{0, 1\}$ such that

$$\xi(g) \leq \epsilon_1 \quad \text{for } g \in G_1, \quad (6.25)$$

$$\sum_{h \in G_1} \delta(h, g) = 1 \quad \text{for } g \in G_1, \quad (6.26)$$

$$\sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \geq \xi(g) \quad \text{for } g, h \in G_1, \quad (6.27)$$

$$\sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \leq \xi(g) + (1 - \delta(h, g))M \quad \text{for } g, h \in G_1, \quad (6.28)$$

where $M = \max_{i \in \{1, \dots, n\}} \max_{g, h \in G_1} D(T_i g \| T_i h)$.

Proof. ($\Rightarrow$) Suppose the distribution $\vec{q}$ on $\{1, \dots, n\}$ satisfies (6.24). Define $\xi : G_1 \rightarrow \mathbb{R}$ and $\delta : G_1 \times G_1 \rightarrow \{0, 1\}$ such that

$$\begin{aligned} \xi(g) &= \min_{h \in G_1} \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \quad \text{for } g \in G_1, \\ \delta(h, g) &= \begin{cases} 1 & \text{if } h = \arg \min \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h), \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

We can check that ξ and δ defined satisfies (6.25) to (6.28).

($\Leftarrow$) Now suppose that there exist functions $\xi : G_1 \rightarrow \mathbb{R}$ and $\delta : G_1 \times G_1 \rightarrow \{0, 1\}$ such that the distribution $\vec{q}$ on $\{1, \dots, n\}$ satisfies (6.25) to (6.28). Fix $g \in G_1$. Since $\sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \geq \xi(g)$, we have $\min_{h \in G_1} \sum_{i=1}^n \vec{q}(i) D(T_i g \| T_i h) \geq \xi(g)$.

Also, since $\sum_{h \in G_1} \delta(h, g) = 1$, there exists a unique h' such that $\delta(h', g) = 1$. Putting this together with (6.27) and (6.28), we obtain $\min_{h \in G_1} \sum_{i=1}^n \bar{q}(i) D(T_i g \| T_i h) = \xi(g)$. By (6.25), we have $\max_{g \in G_1} \min_{h \in G_1} \sum_{i=1}^n \bar{q}(i) D(T_i g \| T_i h) \leq \epsilon_1$.

The proof is complete. $\square$

By Proposition 6.2, Problem (6.20) is equivalent to the following MILP,

$$\begin{aligned}
& \underset{\bar{q}}{\text{maximize}} && \sum_{i=1}^n \bar{q}(i) \mathbb{E}[D(T_i g_I \| T_i f)] \\
& \text{s. t.} && \sum_{i=1}^n \bar{q}(i) D(T_i g \| T_i h) \geq \epsilon_2 \text{ for } g \in G_2 \text{ and } h \in G, \\
& && \xi(g) \leq \epsilon_1 \quad \text{for } g \in G_1, \\
& && \delta(h, g) \in \{0, 1\} \text{ for } g, h \in G_1, \\
& && \sum_{h \in G_1} \delta(h, g) = 1 \text{ for } g \in G_1, \\
& && \sum_{i=1}^n \bar{q}(i) D(T_i g \| T_i h) \geq \xi(g) \text{ for } g, h \in G_1, \\
& && \sum_{i=1}^n \bar{q}(i) D(T_i g \| T_i h) \leq \xi(g) + (1 - \delta(h, g))M \\
& && \text{for } g, h \in G_1.
\end{aligned} \tag{6.29}$$

A global optimal solution to Problem (6.29) can be obtain using branch-and-bound methods on a linear program solver [129].

6.5 Numerical Experiments

6.5.1 Maximal Leakage Privacy

In this subsection, we present numerical results to illustrate the trade-off between the privacy budget ϵ and the expected KL divergence. We consider the signal model with pre-change distribution $\mathbb{P}^f$, set of possible post-change distribution

FIG. 6.1: Comparison of trade-off between the privacy budget ϵ and the expected KL divergence $\mathbb{E}[D(T_{qg_n} || T_{qf})]$ for the Exact and augmented Lagrangian method.

$G = \{\mathbb{P}^{g_1}, \mathbb{P}^{g_2}, \dots, \mathbb{P}^{g_5}\}$ where $\mathbb{P}^f$ and G are randomly generated, uniform prior p_I on the post-change distributions and $\mathcal{X} = \mathcal{Y} = \{1, 2, \dots, 7\}$. Using algorithms described in Section 6.4.1, we solve the relaxed channel design problem (6.18) exactly and the relaxed channel design problem (6.23) using the augmented Lagrangian method. We compare the trade-off between the privacy budget ϵ and the expected KL divergence in Fig. 6.1. When the privacy budget $\epsilon = \log_2 5$, the privacy constraint $\mathcal{L}_{\max}(I \rightarrow J) \leq \epsilon$ becomes redundant and Problem (6.4) reduces to a standard QCD problem. As the augmented Lagrangian method only guarantees local optimality, the channel obtained by solving Problem (6.23) achieves a lower average KL divergence as compared to the channel obtained by solving Problem (6.18). However, as seen in Fig. 6.2, the augmented Lagrangian method requires significantly lesser compute time compared to the Exact method. It should be noted that the compute time required by the Exact Method is directly proportional to $\{|G|_m\}$. The peak observed in Fig. 6.2 corresponds to $\{|G|_m\}$ achieving its maximum at $m = 3$ when $|G| = 5$.

FIG. 6.2: Comparison of the time taken to solve the relaxed channel design problem for the Exact and augmented Lagrangian method.

6.5.2 Sequential Hypothesis Testing Privacy

In this subsection, we present numerical results to illustrate the trade-off between ϵ_1 the privacy of G_1 in G , ϵ_2 the distinguishability of G_2 in G and the expected KL divergence. We consider the signal model with pre-change distribution $\mathbb{P}^f$, set of possible post-change distribution $G = \{\mathbb{P}^{g_1}, \mathbb{P}^{g_2}, \dots, \mathbb{P}^{g_6}\}$ with $G_1 = \{\mathbb{P}^{g_1}, \mathbb{P}^{g_2}, \mathbb{P}^{g_3}\}$ and $G_2 = \{\mathbb{P}^{g_4}, \mathbb{P}^{g_5}, \mathbb{P}^{g_6}\}$, a uniform prior p_I on the post-change distributions, $\mathcal{X} = \mathcal{Y} = \{1, 2, \dots, 4\}$ and the set of deterministic functions from $\mathcal{X}$ to $\mathcal{Y}$ as the finite set of sanitization channel $\mathcal{C}$. We generate f and G randomly. Using algorithms described in Section 6.4.2, we solve the relaxed channel design problem (6.20) by solving the MILP in Problem (6.29). By the data processing inequality $\mathcal{K}_1(T_q) \leq \max_{\mathbb{P}^g \in G_1} \min_{\mathbb{P}^h \in G_1} D(g \| h) = 0.9387$ and $\mathcal{K}_2(T_q) \leq \min_{\mathbb{P}^g \in G_2} \min_{\mathbb{P}^h \in G} D(g \| h) = 0.0855$, thus we focus our attention to the region where $0 \leq \epsilon_1 \leq 0.9387$ and $0 \leq \epsilon_2 \leq 0.0855$. We compare the trade-off between the privacy ϵ_1 of G_1 in G and the expected KL divergence in Fig. 6.3 and the trade-off between the distinguishability ϵ_2 of G_1 in G and the expected KL divergence in Fig. 6.4. In Fig. 6.4, it can be observed that Problem (6.20) is not feasible for $\epsilon_2 \geq 0.0233$. In both Fig. 6.3 and Fig. 6.4, we observe that the average KL divergence increases as ϵ_1 increases and

ϵ_2 decreases which is consistent with the behaviour expected of the optimal value of Problem (6.20).

FIG. 6.3: Comparison of trade-off between the privacy ϵ_1 of G_1 in G and the expected KL divergence for different distinguishability ϵ_2 of G_2 in G .

FIG. 6.4: Comparison of trade-off between the distinguishability ϵ_2 of G_2 in G and the expected KL divergence for different privacy ϵ_1 of G_1 in G .

Chapter 7

Conclusion and Future Work

In this thesis, we investigated the non-Bayesian QCD problem under various operational constraints. We have developed methods and derived bounds on their performances under each of the operational constraints.

In Chapter 3, we have studied the sequential change detection problem when the pre-change distribution $\mathbb{P}^f$ is known, and the post-change distribution $\mathbb{P}^g$ lies in a family of distributions with k -th moment differing from $\mathbb{P}^f$ by at least ϵ . We proposed a sequential change detection method that partitions the sample space into bins and a test statistic that can be updated recursively in real-time with low complexity. We analyzed the growth rate of our test statistic and used it to heuristically deduce the asymptotic relationship between the ADD and the ARL. Tests on both synthetic and real data suggest that our proposed BG-CuSum test outperforms several other non-parametric tests in the literature. Furthermore, simulations indicate that the BG-CuSum test approaches the performance of $\hat{\tau}$, which has known asymptotic properties, as the ARL becomes large. We provided a lower bound on the ARL of the BG-CuSum test to aid the setting of the threshold b . One direction for future work would be to derive the WADD for the BG-CuSum test. This remains an open research problem due to the technical difficulties introduced by having to estimate the post-change distribution using an estimated

change-point. Although we have assumed that the pre- and post-change distributions are defined on $\mathbb{R}$, the BG-CuSum test derived in this chapter can be applied to cases where f and g are generalized pdfs on $\mathbb{R}^n$ with $n > 1$. To see how this can be done, we assume that $g \in D(f, N)$, where $N = N_0^n$ is a power of n . We can then divide $\mathbb{R}^n$ into N equi-probable sets w.r.t. f_c by sequentially dividing each dimension into N_0 equi-probable intervals. After obtaining these N sets, we can apply the BG-CuSum test directly. The results derived in Section 3.3 can also be extended to the case where the distributions are on $\mathbb{R}^n$. However, the amount of data required to learn the pre-change distribution $\mathbb{P}^f$ increases quickly w.r.t. n , which limits its application in practice. The assumption that the post-change distribution $\mathbb{P}^g$ is absolutely continuous w.r.t. $\mathbb{P}^f$ can be further relaxed. For example, the results developed in Section 3.2 and 3.3 extend to the case with a continuous pre-change distribution $\mathbb{P}^f$ and a discrete post-change distribution $\mathbb{P}^g$. The recursive update scheme in Theorem 3.1 and the lower bound on the ARL in Proposition 3.1 still holds. However, as $\mathbb{P}^g$ is not absolutely continuous w.r.t. $\mathbb{P}^f$, the instantaneous empirical log-likelihood ratio $\log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)} = \infty$ and the instantaneous log-likelihood ratio $\log \frac{\hat{g}_N(X_i)}{f_N(X_i)} = \infty$ with positive probability under $\mathbb{P}_1$. Thus, we have that $D(g_N \| f_N) = \infty$. Furthermore, in Proposition 3.4, the empirical average $\frac{1}{t} \sum_{i=1}^t \log \frac{\hat{g}_N^{1:i-1}(X_i)}{f_N(X_i)}$ diverges to infinity almost surely.

In Chapter 4, we have studied the non-Bayesian QCD problem where the signal may be subjected to a nuisance change. We proposed the W-SGLR stopping time that quickly detects the critical change while ignoring the nuisance change. The limited window size ensures that the W-SGLR stopping time does not require increasing computational resources as more samples are observed. We also derived the stopping time's asymptotic behavior and showed that it is asymptotically optimal under mild technical assumptions. A generalized W-SGLR stopping time is also proposed for the case where the critical and nuisance post-change distributions are unknown but belong to a parametrized family. Numerical simulations and experiments on a real dataset demonstrated that the W-SGLR stopping time achieves better ADD-ARL trade-off than various other competing stopping times. In this

chapter, we have assumed that if both the critical and nuisance changes occur, the eventual distribution that generates the signal is the same, regardless of which change comes first. A more general model would be to allow the eventual distribution to depend on the order of the change points. An easy generalization of the W-SGLR stopping would be to include all the different eventual distributions into the numerator of (4.13). The asymptotic trade-off between the WADD and ARL can be derived using similar techniques in Section 4.2.1. However, deriving the conditions for asymptotic optimality of this stopping time is more complicated and would be a possible direction for future research. Another possible future research direction is to consider a modification of the W-SGLR stopping time for the TCD problem under the possibility of a nuisance change. As the performance metrics of the TCD problem are different from the QCD problem, its asymptotic trade-off between the worst-case false alarms and missed detection within a specified window needs to be studied.

In Chapter 5, we discussed the problem of QCD while taking sampling and switching costs into consideration. We formulated the QCD problem with an additional AOSC constraint. Asymptotically optimal stopping times were proposed and the design of optimal sampling policies was formulated as quadratic optimization problems. We showed that sampling policies of window size $W > 1$ can be reduced to a sampling policy of window size $W = 1$ while maintaining the same asymptotic ARL-WADD trade-off and AOSC. Thus, it is sufficient to solve the policy design problem for $W = 0, 1$. We applied the IRM algorithm to the policy design problem to obtain locally optimal solutions. For cases with additional assumptions on the observation-switching cost matrix $\mathcal{C}$, globally optimal solutions can be obtained. The methods developed are for the case when the window W is finite. The results regarding the structure of asymptotically optimal stopping times such as Lemma 5.1, Proposition 5.2 and Theorem 5.1, do not hold in general when $W = \infty$ and would be an interesting direction for future work. In some Internet of Things (IoT) applications, the nodes collecting observations have some computing resources to perform local computations and observations. In these applications,

the individual sensors can make use of aggregated information to decide whether it is necessary to sample while taking into account sampling and communication costs. This would require us to consider active rather than static sampling policies for QCD and would be an interesting area for further research.

In Chapter 6, we have proposed a framework for privacy-aware QCD using two different privacy metrics. We also proposed optimization problems for which the solution provides a sanitization channel where the GLR CuSum stopping time is asymptotically optimal. We derived relaxations to the channel design problem and provided algorithms to obtain an exact solution to the relaxed channel design problem. For the maximal leakage privacy metric, a continuous relaxation for the channel design problem is proposed so that locally optimal solutions can be obtained when the exact solutions are computationally intractable. One drawback of the proposed solution for privacy-aware QCD is the discreteness of the privacy constraint $\mathcal{L}_{\max}(I \rightarrow J)$. This may lead to some inefficiencies in channel design. Using multiple possible sanitization channels, the privacy constraint $\mathcal{L}_{\max}(I \rightarrow J)$ can achieve a continuous interval of values. The optimization problem related to the new channel design problem is, however, much harder to solve and is a potential direction for future work. For the sequential hypothesis testing privacy metric, a modified observation model is proposed to relax the channel design problem into a MILP. One interesting direction of future work can be done to determine the pairs (ϵ_1, ϵ_2) where Problem (6.20) is feasible.

In this thesis, we have considered only signal models where the observations are independent given the change-points. An interesting direction for future research is to consider the more general signal model, such as hidden Markov models, where the observations given the change-points have dependencies.

List of Author's Publications

The following is a list of publications in refereed journals and conference proceedings produced during my PhD candidature. In some cases, the journal papers contain partial material overlapping with the conference publications.

Journal Articles

- **T. S. Lau** and W. P. Tay, “Quickest change detection in the presence of a nuisance change,” *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5281 – 5296, Oct. 2019.
- **T. S. Lau**, W. P. Tay, and V. V. Veeravalli, “A binning approach to quickest change detection with unknown post-change distribution,” *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 609 – 621, Feb. 2019.

Conference Proceedings

- **T. S. Lau** and W. P. Tay, “Asymptotically optimal quickest change detection under a nuisance change,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Brighton, UK, May 2019*.
- **T. S. Lau** and W. P. Tay, “Quickest change detection under a nuisance change,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Calgary, Canada, Apr. 2018*.

- **T. S. Lau** and W. P. Tay, “Optimal sampling policy for quickest change detection,” in *Proc. IEEE Global Conf. on Signal and Information Processing, Montreal, Canada, Nov. 2017*.
- **T. S. Lau**, W. P. Tay, and V. V. Veeravalli, “Quickest change detection with unknown post-change distribution,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, New Orleans, USA, Mar. 2017*.

Conference and Journal Articles Under Submission/Revision

- **T. S. Lau** and W. P. Tay, “Privacy-aware quickest change detection,” submitted to *the IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Barcelona, Spain, May 2020*.
- **T. S. Lau**, W. P. Tay, “Asymptotically optimal sampling policy for quickest change detection with observation-switching cost,” submitted to *IEEE Trans. Signal Process.*.

Bibliography

- [1] A. Shiryaev, “The problem of the most rapid detection of a disturbance in a stationary process,” in *Soviet Math. Dokl.*, vol. 2, no. 795-799, 1961. [1](#)
- [2] A. N. Shiryaev, “On optimum methods in quickest detection problems,” *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963. [1](#)
- [3] ———, *Optimal stopping rules*. Springer Science & Business Media, 2007, vol. 8. [1](#)
- [4] A. G. Tartakovsky and V. V. Veeravalli, “Asymptotic analysis of Bayesian quickest change detection procedures,” in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2002, p. 217. [1](#)
- [5] ———, “General asymptotic Bayesian theory of quickest change detection,” *Theory of Probability & Its Applications*, vol. 49, no. 3, pp. 458–497, 2005. [1](#), [6](#)
- [6] E. S. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954. [2](#), [20](#), [26](#), [67](#)
- [7] G. Lorden, “Procedures for reacting to a change in distribution,” *Ann. Math. Stat.*, pp. 1897–1908, 1971. [2](#), [20](#), [25](#), [65](#), [66](#), [68](#), [111](#), [145](#)
- [8] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *Ann. Stat.*, pp. 1379–1387, 1986. [2](#)
- [9] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, 1998. [2](#), [4](#), [69](#), [75](#), [79](#), [106](#)
- [10] M. Pollak, “Optimal detection of a change in distribution,” *Ann. Stat.*, pp. 206–227, 1985. [2](#)
- [11] D. Siegmund and E. S. Venkatraman, “Using the generalized likelihood ratio statistic for sequential detection of a change-point,” *Ann. Stat.*, pp. 255–271, 1995. [2](#), [4](#)
- [12] T. Banerjee and V. V. Veeravalli, “Data-efficient minimax quickest change detection with composite post-change distribution,” *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 5172–5184, 2015. [2](#), [4](#), [8](#)

- [13] T. S. Lau, W. P. Tay, and V. V. Veeravalli, “Quickest change detection with unknown post-change distribution,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.* IEEE, 2017, pp. 3924–3928. [2](#)
- [14] T. S. Lau and W. P. Tay, “Optimal sampling policy for quickest change detection,” in *Proc. IEEE Global Conf. Signal and Inform. Process.* [2](#), [8](#), [145](#), [150](#)
- [15] T. S. Lau, W. P. Tay, and V. V. Veeravalli, “A binning approach to quickest change detection with unknown post-change distribution,” *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 609–621, Feb 2019. [2](#)
- [16] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014. [2](#), [15](#), [30](#), [51](#), [116](#), [143](#), [145](#)
- [17] V. V. Veeravalli and T. Banerjee, “Quickest change detection,” *Academic press library in signal process.: Array and statistical signal process.*, vol. 3, pp. 209–256, 2013. [2](#)
- [18] H. V. Poor and O. Hadjiladis, *Quickest detection*. Cambridge University Press Cambridge, 2009, vol. 40. [2](#), [15](#), [51](#), [68](#), [74](#), [80](#), [102](#), [105](#), [116](#), [119](#)
- [19] W. H. Woodall, D. J. Spitzner, D. C. Montgomery, and S. Gupta, “Using control charts to monitor process and product quality profiles,” *J. of Quality Technology*, vol. 36, no. 3, p. 309, 2004. [2](#)
- [20] T. L. Lai, “Sequential changepoint detection in quality control and dynamical systems,” *J. of the Roy. Statistical Soc.*, pp. 613–658, 1995. [2](#)
- [21] H. Sohn, J. A. Czarnecki, and C. R. Farrar, “Structural health monitoring using statistical process control,” *J. Structural Eng.*, vol. 126, no. 11, pp. 1356–1363, 2000. [2](#)
- [22] S. Xie, G. Wang, S. Lin, and P. S. Yu, “Review spam detection via temporal pattern discovery,” in *Proc. Conf. Knowl. Discovery and Data Mining*. ACM, 2012, pp. 823–831. [2](#)
- [23] F. Ji, W. P. Tay, and L. Varshney, “An algorithmic framework for estimating rumor sources with different start times,” vol. 65, no. 10, pp. 2517 – 2530, May 2017. [2](#)
- [24] W. Tang, F. Ji, and W. P. Tay, “Estimating infection sources in networks using partial timestamps,” vol. 13, no. 12, pp. 3035–3049, Dec. 2018. [2](#)
- [25] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” *Statistical Sci.*, pp. 235–249, 2002. [2](#)
- [26] L. Lai, Y. Fan, and H. V. Poor, “Quickest detection in cognitive radio: A sequential change detection framework,” in *IEEE Conf. Global Telecommun.* IEEE, 2008, pp. 1–5. [2](#)

- [27] V. M. R. Muggeo and G. Adelfio, “Efficient change point detection for genomic sequences of continuous measurements,” *Bioinformatics*, vol. 27, no. 2, pp. 161–166, 2011. [2](#)
- [28] T. Banerjee, Y. C. Chen, A. D. Dominguez-Garcia, and V. V. Veeravalli, “Power system line outage detection and identification—a quickest change detection approach,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.* IEEE, 2014, pp. 3450–3454. [2](#)
- [29] P. R. Coppin and M. E. Bauer, “Digital change detection in forest ecosystems with remote sensing imagery,” *Remote sensing reviews*, vol. 13, no. 3-4, pp. 207–234, 1996. [2](#)
- [30] Y.-W. Hong and A. Scaglione, “Distributed change detection in large scale sensor networks through the synchronization of pulse-coupled oscillators,” vol. 3, May 2004, pp. iii–869. [2](#)
- [31] J. Yang, X. Zhong, and W. P. Tay, “A dynamic Bayesian nonparametric model for blind calibration of sensor networks,” *IEEE Internet of Things J.*, vol. 5, no. 5, pp. 3942–3953, Oct. 2018. [2](#)
- [32] K. Sequeira and M. Zaki, “ADMIT: anomaly-based data mining for intrusions,” in *Proc. Conf. Knowl. Discovery and Data Mining.* ACM, 2002, pp. 386–395. [2](#)
- [33] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blazek, and H. Kim, “A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods,” *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3372–3382, 2006. [2](#)
- [34] W. Luo, W. P. Tay, and M. Leng, “Infection spreading and source identification: A hide and seek game,” vol. 64, no. 16, pp. 4228 – 4243, Aug. 2016. [2](#)
- [35] W. A. Smith and R. B. Randall, “Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study,” *Mechanical Syst. and Signal Process.*, vol. 64, pp. 100–131, 2015. [2](#), [5](#), [64](#), [80](#), [90](#)
- [36] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011. [2](#), [31](#), [38](#)
- [37] M. D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *J. Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 577–588, 1995. [3](#)
- [38] J. Q. Li and A. R. Barron, “Mixture density estimation.” in *Proc. Advances in Neural Inform. Process. Syst.*, 1999, pp. 279–285. [3](#)
- [39] M. Pawlak and A. Steland, “Nonparametric sequential signal change detection under dependent noise,” *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3514–3531, June 2013. [4](#)

- [40] E. Sakurai and K. Yamanishi, “Comparison of dynamic model selection with infinite HMM for statistical model change detection,” in *Proc. IEEE Inform. Theory Workshop*, Sept 2012, pp. 302–306. 4
- [41] L. Gordon and M. Pollak, “An efficient sequential nonparametric scheme for detecting a change of distribution,” *Ann. Stat.*, pp. 763–804, 1994. 4
- [42] D. M. Hawkins and Q. Deng, “A nonparametric change-point control chart,” *J. Quality Technology*, vol. 42, no. 2, 2010. 4, 5, 31, 32, 33, 40
- [43] E. Rafajlowicz, M. Pawlak, and A. Steland, “Nonparametric sequential change-point detection by a vertically trimmed box method,” *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3621–3634, July 2010. 4
- [44] ———, “Nonparametric sequential change-point detection by a vertical regression method,” in *Proc. IEEE Workshop on Statistical Signal Process.*, Aug 2009, pp. 614–617. 4
- [45] M. Lang and A. M. Zoubir, “A nonparametric cumulative sum scheme based on sequential ranks and adaptive control limits,” in *Proc. European Signal Process. Conf.*, Aug 2015, pp. 1984–1988. 4
- [46] G. J. Ross, D. K. Tasoulis, and N. M. Adams, “Nonparametric monitoring of data streams for changes in location and scale,” *Technometrics*, vol. 53, no. 4, pp. 379–389, 2011. 4
- [47] S. Li, Y. Xie, H. Dai, and L. Song, “M-statistic for kernel change-point detection,” in *Proc. Advances Neural Inform. Process. Syst.*, 2015, pp. 3366–3374. 4, 5
- [48] B. Darkhovskii and B. Brodskii, “A nonparametric method for fastest detection of a change in the mean of a random sequence,” *Theory of Probability & Its Applicat.*, vol. 32, no. 4, pp. 640–648, 1988. 4, 32
- [49] S. Nitinawarat and V. V. Veeravalli, “Universal quickest outlier detection and isolation,” in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2015, pp. 770–774. 5
- [50] G. J. Ross and N. M. Adams, “Two nonparametric control charts for detecting arbitrary distribution changes,” *J. Quality Technology*, vol. 44, no. 2, p. 102, 2012. 5, 32, 33, 40
- [51] C.-D. Fuh, “SPRT and CuSum in hidden Markov models,” *Ann. Statist.*, vol. 31, no. 3, pp. 942–977, 06 2003. [Online]. Available: <https://doi.org/10.1214/aos/1056562468> 6
- [52] ———, “Asymptotic operating characteristics of an optimal change point detection in hidden Markov models,” *Annals of Statistics*, pp. 2305–2339, 2004. 6

- [53] C.-D. Fuh and Y. Mei, “Quickest change detection and Kullback-Leibler divergence for two-state hidden Markov models,” *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4866–4878, 2015. [6](#)
- [54] G. V. Moustakides, “Quickest detection of abrupt changes for a class of random processes,” *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1965–1968, 1998. [6](#)
- [55] V. Krishnamurthy, “Quickest detection pomdps with social learning: Interaction of local and global decision makers,” *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5563–5587, Aug 2012. [6](#)
- [56] B. K. Guépié, L. Fillatre, and I. Nikiforov, “Sequential detection of transient changes,” *Sequential Analysis*, vol. 31, no. 4, pp. 528–547, 2012. [6](#), [66](#)
- [57] E. Ebrahimzadeh and A. Tchamkerten, “Sequential detection of transient changes in stochastic systems under a sampling constraint,” in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2015, pp. 156–160. [6](#)
- [58] G. V. Moustakides and V. V. Veeravalli, “Sequentially detecting transitory changes,” in *Proc. IEEE Int. Symp. Inform. Theory*, July 2016, pp. 26–30. [6](#)
- [59] S. Zou, G. Fellouris, and V. V. Veeravalli, “Quickest change detection under transient dynamics: Theory and asymptotic analysis,” *IEEE Trans. Inf. Theory*, 2018. [6](#)
- [60] J. Heydari and A. Tajer, “Quickest search for a change point,” *arXiv preprint arXiv:1804.09657*, 2018. [6](#)
- [61] X. Liu, W. P. Tay, Z.-W. Liu, and G. Xiao, “Quasi-synchronization of heterogeneous networks with a generalized Markovian topology and event-triggered communication,” 2020, in press. [7](#)
- [62] F. Ji and W. P. Tay, “A Hilbert space theory of generalized graph signal processing,” vol. 67, no. 24, pp. 6188 – 6203, Dec. 2019. [7](#)
- [63] V. V. Veeravalli, T. BaSar, and H. V. Poor, “Decentralized sequential detection with a fusion center performing the sequential test,” *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 433–442, 1993. [7](#)
- [64] V. V. Veeravalli, “Decentralized quickest change detection,” *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1657–1665, 2001. [7](#)
- [65] B. Jiang and L. Lai, “Distributed quickest detection with optional observations at the fusion center,” in *IEEE Conf. Signals, Syst. and Comput.*, IEEE, 2016, pp. 112–116. [7](#)
- [66] O. Hadjiliadis, H. Zhang, and H. V. Poor, “One shot schemes for decentralized quickest change detection,” *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3346–3359, 2009. [7](#)

- [67] T. Banerjee and V. V. Veeravalli, “Data-efficient quickest change detection with on–off observation control,” *Sequential Anal.*, vol. 31, no. 1, pp. 40–77, 2012. [8](#)
- [68] —, “Data-efficient quickest change detection in minimax settings,” *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6917–6931, 2013. [8](#), [111](#)
- [69] T. Banerjee, V. V. Veeravalli, and A. Tartakovsky, “Decentralized data-efficient quickest change detection,” in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2013, pp. 2587–2591. [8](#)
- [70] Y. Xie, M. Wang, and A. Thompson, “Sketching for sequential change-point detection,” in *Proc. IEEE Global Conf. Signal and Inform. Process.* IEEE, 2015, pp. 78–82. [8](#)
- [71] G. K. Atia, “Change detection with compressive measurements,” *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 182–186, 2015. [8](#)
- [72] J. Heydari and A. Tajer, “Quickest change detection in structured data with incomplete information,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, March 2017, pp. 6434–6438. [8](#)
- [73] 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. [8](#)
- [74] M. Sun and W. P. Tay, “Privacy-preserving nonparametric decentralized detection,” Shanghai, China, Mar. 2016. [10](#)
- [75] X. He and W. P. Tay, “Multilayer sensor network for information privacy,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2017. [10](#)
- [76] M. Sun and W. P. Tay, “Inference and data privacy in IoT networks,” 2017. [10](#)
- [77] —, “On the relationship between inference and data privacy in decentralized iot networks,” *IEEE Trans. Inf. Forensics Security*, 2019. [10](#)
- [78] Y. Wang, X. Wu, and H. Donghui, “Using randomized response for differential privacy preserving data collection,” in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Washington, D.C.*, 2003, pp. 505–510. [10](#)
- [79] A. D. Sarwate and L. Sankar, “A rate-distortion perspective on local differential privacy,” in *Proc. Allerton Conf. on Commun., Control and Computing, Monticello, IL*, 2014, pp. 903–908. [10](#)
- [80] S. Xiong, A. D. Sarwate, and N. B. Mandayam, “Randomized requantization with local differential privacy,” Shanghai, 2016, pp. 2189–2193. [10](#)

- [81] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in *Proc. IEEE Symp. on Foundations of Computer Science*, Berkeley, 2013, pp. 429–438. [10](#)
- [82] R. Gupta and U. P. Rao, “Achieving location privacy through CAST in location based services,” *IEEE J. Commun. Netw.*, vol. 19, no. 3, pp. 239–249, Jul. 2017. [10](#)
- [83] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-DNF formulas on ciphertexts,” in *Proc. Int. Conf. on Theory of Cryptography*, Cambridge, MA, 2005, pp. 325–341. [10](#)
- [84] Y. Ishai and A. Paskin, “Evaluating branching programs on encrypted data,” in *Proc. Int. Conf. on Theory of Cryptography*, Berlin, Heidelberg, 2007, pp. 575–594. [10](#)
- [85] C. Gentry, “Fully homomorphic encryption using ideal lattices.” Bethesda, MD, 2009, pp. 169–178. [10](#)
- [86] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, “How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data,” Austin, TX, 2013, pp. 269–272. [10](#)
- [87] H. Yamamoto, “A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers,” vol. 29, no. 6, pp. 918–923, 1983. [10](#)
- [88] M. Khouzani and P. Malacaria, “Generalized entropies and metric-invariant optimal countermeasures for information leakage under symmetric constraints,” *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 888–901, Feb 2019. [10](#)
- [89] W. Wang, L. Ying, and J. Zhang, “On the relation between identifiability, differential privacy, and mutual-information privacy,” *IEEE Trans. Inf. Theory*, vol. 62, pp. 5018–5029, Sep 2016. [10](#)
- [90] H. Song, G. A. Fink, and S. Jeschke, *Tutorial on Information Theoretic Metrics Quantifying Privacy in Cyber-Physical Systems*. IEEE, 2017. [Online]. Available: <https://ieeexplore.ieee.org/document/8068891> [10](#)
- [91] I. Issa, S. Kamath, and A. B. Wagner, “An operational measure of information leakage,” in *Proc. Ann. Conf. on Inf. Sci. and Syst.*, March 2016, pp. 234–239. [10](#), [142](#), [148](#)
- [92] I. Issa and A. B. Wagner, “Operational definitions for some common information leakage metrics.” IEEE, 2017, pp. 769–773. [10](#)
- [93] C. Dwork, *Differential Privacy*, 2016. [10](#)

- [94] Z. Li, T. J. Oechtering, and D. s, "Privacy against a hypothesis testing adversary," *IEEE Trans. on Information Forensics and Security*, vol. 14, no. 6, pp. 1567–1581, June 2019. [10](#)
- [95] Y. Song, C. X. Wang, and W. P. Tay, "Compressive privacy for a linear dynamical system," *IEEE Trans. Inf. Forensics Security*, 2019. [10](#)
- [96] S. Y. Kung, "Compressive privacy from information estimation," *IEEE Signal Process. Mag.*, vol. 34, no. 1, pp. 94–112, Jan 2017. [10](#)
- [97] S. Asoodeh, F. Alajaji, and T. Linder, "Privacy-aware MMSE estimation," Barcelona, Spain, Jul. 2016. [10](#)
- [98] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," Monticello, IL, Oct. 2012. [10](#)
- [99] X. He and W. P. Tay, "Multilayer sensor network for information privacy," New Orleans, USA, Mar. 2017. [10](#)
- [100] M. Sun, W. P. Tay, and X. He, "Toward information privacy for the internet of things: A nonparametric learning approach," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1734–1747, Dec 2018. [10](#)
- [101] M. Sun and W. P. Tay, "Decentralized detection with robust information privacy protection," *IEEE Trans. Inf. Forensics Security*, vol. 15, no. 1, pp. 85–99, Nov. 2019. [10](#), [11](#)
- [102] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 594–603, 2020. [11](#)
- [103] J. Liao, L. Sankar, F. P. Calmon, and V. Y. Tan, "Hypothesis testing under maximal leakage privacy constraints," Aachen, Germany, 2017, pp. 779–783. [11](#)
- [104] R. Cummings, S. Krehbiel, Y. Mei, R. Tuo, and W. Zhang, "Differentially private change-point detection," in *Advances in Neural Inform. Process. Syst.*, 2018, pp. 10 825–10 834. [11](#)
- [105] R. Durrett, *Probability: Theory and examples*. Cambridge University Press, 2010. [15](#)
- [106] W. Rudin, *Real and complex analysis*. Tata McGraw-Hill Education, 1987. [17](#), [24](#), [57](#), [60](#)
- [107] A. Wald, "Sequential tests of statistical hypotheses," *The Ann. of Math. Stat.*, vol. 16, no. 2, pp. 117–186, 1945. [17](#)
- [108] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Ann. of Math. Stat.*, pp. 326–339, 1948. [19](#)

- [109] 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. [25](#)
- [110] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 1985. [43](#)
- [111] R. N. Bhattacharya and E. C. Waymire, *A basic course in probability theory*. Springer, 2007, vol. 69. [44](#)
- [112] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *J. of the Amer. statistical association*, vol. 58, no. 301, pp. 13–30, 1963. [45](#)
- [113] Y. Mei, “Asymptotically optimal methods for sequential change-point detection,” Ph.D. dissertation, California Institute of Technology, 2003. [65](#)
- [114] T. Bojdecki, “Probability maximizing approach to optimal stopping and its application to a disorder problem,” *Stochastics*, vol. 3, no. 1-4, pp. 61–71, 1980. [66](#)
- [115] M. Pollak and A. M. Krieger, “Shewhart revisited,” *Sequential Analysis*, vol. 32, no. 2, pp. 230–242, 2013. [66](#)
- [116] G. V. Moustakides, “Multiple optimality properties of the Shewhart test,” *Sequential Analysis*, vol. 33, no. 3, pp. 318–344, 2014. [66](#)
- [117] D. Egea-Roca, J. A. López-Salcedo, G. Seco-Granados, and H. Vincent Poor, “Performance bounds for finite moving average tests in transient change detection,” *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1594–1606, March 2018. [66](#), [83](#)
- [118] A. Willsky and H. Jones, “A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems,” *IEEE Trans. Autom. Control*, vol. 21, no. 1, pp. 108–112, 1976. [69](#), [79](#)
- [119] L. Lai, Y. Fan, and H. V. Poor, “Quickest detection in cognitive radio: A sequential change detection framework,” in *Proc. IEEE Global Telecomm. Conf.*, Nov 2008, pp. 1–5. [74](#), [87](#)
- [120] Y. Wang, A. Yang, X. Chen, P. Wang, Y. Wang, and H. Yang, “A deep learning approach for blind drift calibration of sensor networks,” *IEEE Sensors J.*, vol. 17, no. 13, pp. 4158–4171, July 2017. [74](#)
- [121] Z. Li, Y. Wang, A. Yang, and H. Yang, “Drift detection and calibration of sensor networks,” in *Proc. Int. Conf. on Wireless Commun. Signal Process.*, Oct 2015, pp. 1–6. [74](#)
- [122] M. Takruri and S. Challa, “Drift aware wireless sensor networks,” in *Proc. Int. Conf. on Inform. Fusion*, July 2007, pp. 1–7. [74](#)

- [123] T. S. Lau and W. P. Tay, “Quickest change detection in the presence of a nuisance change,” *Arxiv*, 2019. [79](#), [96](#)
- [124] ———, “Quickest change detection under a nuisance change,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.* IEEE, 2018, pp. 6643–6647. [83](#)
- [125] ———, “Asymptotically optimal quickest change detection under a nuisance change,” 2019. [93](#)
- [126] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004. [105](#)
- [127] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*. Athena Scientific Belmont, MA, 2002, vol. 1. [113](#)
- [128] D. H. Greene and D. E. Knuth, *Mathematics for the Analysis of Algorithms*. Springer Science & Business Media, 2007. [116](#)
- [129] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” <http://cvxr.com/cvx>, Mar. 2014. [121](#), [126](#), [154](#)
- [130] ———, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html. [121](#)
- [131] C. Sun and R. Dai, “Rank-constrained optimization and its applications,” *Automatica*, vol. 82, pp. 128–136, 2017. [121](#), [122](#), [123](#)
- [132] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, “Learning laplacian matrix in smooth graph signal representations,” *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, 2016. [126](#)
- [133] J. Riordan, *Introduction to combinatorial analysis*. Courier Corporation, 2012. [151](#)
- [134] R. E. Perez, P. W. Jansen, and J. R. R. A. Martins, “pyOpt: A Python-based object-oriented framework for nonlinear constrained optimization,” *Structures and Multidisciplinary Optimization*, vol. 45, no. 1, pp. 101–118, 2012. [152](#)
- [135] A. R. Conn, G. Gould, and P. L. Toint, *LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A)*. Springer Science & Business Media, 2013, vol. 17. [152](#)
- [136] D. Fernández and M. V. Solodov, “Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition,” *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 384–407, 2012. [152](#)