

國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

結構變動之易辛模型的最快速變化偵測

Quickest Change Detection
for Structural Changing Ising Model

林家緯

Jia-wei Lin

指導教授: 王奕翔 博士

Advisor: I-Hsiang Wang Ph.D.

中華民國 110 年 11 月

November, 2021





國立臺灣大學（碩）博士學位論文
口試委員會審定書

結構變動之易辛模型的最快速變化偵測
Quickest Change Detection

for Structural Changing Ising Model

本論文係林家緯君（r08942058）在國立臺灣大學電信工程學研究所完成之碩（博）士學位論文，於民國 110 年 11 月 29 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

王炎翔

（簽名）

（指導教授）

黃昱智

林士駿

所 長

劉錫坤

（簽名）





Acknowledgements

很感謝王奕翔教授這兩年的指導，讓我得以真正學習做研究。從一開始不清楚如何讀論文、對於研究沒有自己的想法或是陷於太複雜的問題情境，到逐步能夠主動發問並建立自己的看法，甚至偶而能夠從直覺下手進而有量化的結果。我認為這套思維上的訓練不僅僅能用在研究上，也能對每個學習的機會有正面幫助。此外，每週的個人或是團體會議也是很好的練習口語表達的機會，很感謝老師的指點。

同時我也感謝實驗室的眾多同學在這段期間的討論與鼓勵，讓我能夠更快掌握做研究的一些思維，也使我學習以不同的角度來看待問題。

最後，特別感謝家人的支持，讓我能無後顧之憂地完成學業。





摘要

本篇碩士論文在探討具有相關性的高維機率分布發生結構變異的最快速變化偵測問題。我們假設網路上蒐集的資料是基於易辛模型且在發生變化前後，資料的機率分布唯一的差異是易辛模型中的結構。在這樣的假設底下，我們發現此問題與伯努利的最快速變化偵測問題高度相關，甚至當我們適當強化易辛模型的假設，相應的伯努利最快速變化偵測問題將被大大簡化。正因如此，我們能夠提出了一個在新增多條邊的簡化易辛模型問題上，只需要知道變化前網路結構，就能夠達到最佳的最壞平均檢測延遲與平均時間至錯誤警告平衡的方法。在減少一條邊的簡化易辛模型問題上，我們雖然沒能夠在只知曉變化前網路結構下，做到最佳的最壞平均檢測延遲與平均時間至錯誤警告平衡，但是我們的方法從文獻中另一個評量標標準來看，已經達到最優。由於上述提出的方法在實際執行上複雜度較高，我們進一步利用了易辛模型的關聯性傳播特性，提出只需從網路上蒐集少數幾個節點的資料，就能夠順利偵測變化的方法。由於節點的選擇將會影響偵測的表現，我們將選擇解點的問題寫成一個最佳化問題，並提出解決此問題的演算法。

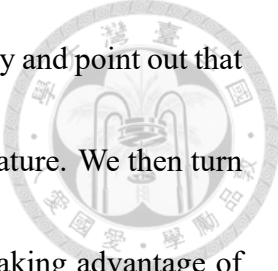
關鍵字：最快速變化偵測、易辛模型





Abstract

The problem of quickest change detection for structural changes within correlated, structured distribution is studied. Precisely, we assume that data collected over a network follows a specific instance of undirected graphical models, the Ising model and that the Ising model before and after the change differs only in their structures. We show that for both types of structural change, edge appearance and disappearance, the problem can be suitably transformed into that of Bernoulli random variables. When further restricting the Ising model to have zero mean-field vectors and forest interaction structure, the transformed Bernoulli problem becomes drastically simpler to solve. Consequently, under the multiple edge appearance setting in Ising forests, our proposed algorithm equipped with only the pre-change structure knowledge is capable of achieving the optimal worst average detection delay and average run length to false alarm trade-off at the large average run length to false alarm regime. Under the one edge disappearance setting in Ising forests, we present an algorithm with sub-optimal worst average detection delay and average run



length to false alarm trade-off. We explain the reason for sub-optimality and point out that our method is already optimal under a less strict criteria from the literature. We then turn to improve the runtime complexity of the proposed algorithms. By taking advantage of the correlation propagation nature of correlated networks, we show that instead of doing brute force search for all possible locations of the structural change, monitoring only a few chosen nodes in the network suffice for decent statistical performance. To get the most out of the new procedure, we formulate a node selection optimization problem and provide a simple algorithm to solve it.

Keywords: Quickest change detection, Ising models



Contents

	Page
Verification Letter from the Oral Examination Committee	i
Acknowledgements	iii
摘要	v
Abstract	vii
Contents	ix
List of Figures	xiii
Denotation	xv
Chapter 1 Introduction	1
1.1 Contribution of Thesis	5
Chapter 2 Backgrounds	9
2.1 Quickest Change Detection (QCD)	9
2.1.1 Composite QCD	13
2.1.1.1 Post-change Unknown QCD	14
2.1.1.2 Pre-change Unknown QCD	15
2.1.1.3 Pre/Post-change Unknown QCD	16
2.2 Undirected Graphical Models	17
2.2.1 Ising Models	19
2.2.1.1 Ising Model in a Forest	20

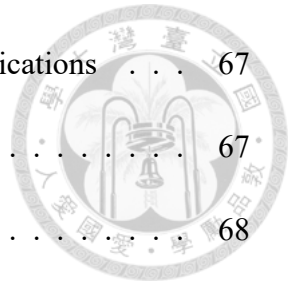
Chapter 3 QCD for Edge Increased in Ising Models on a Forest

23

3.1	One Edge Increased/One Component Decreased with Known Pre-Change Distribution	24
3.1.1	Problem Setting	24
3.1.2	Proposed Algorithm	24
3.2	One Edge Increased/One Component Decreased with only Structure of Pre-Change Distribution	26
3.2.1	Problem Setting	26
3.2.2	Proposed Algorithm	26
3.2.3	Important Property	27
3.2.4	Performance of Proposed Algorithm	28
3.3	Multiple Edge Increased/One Component Decreased with only Structure of Pre-Change Distribution	30
3.3.1	Problem Setting	30
3.3.2	Proposed Algorithm	30
3.3.3	Performance of Proposed Algorithm	31
3.4	Representative Procedure for Edge Increased/Component Decreased .	34
3.4.1	Proposed Algorithm	35
3.4.2	Performance for Proposed Algorithm	36
3.4.3	Choosing the Representatives	38
3.5	Alternatives to Generalized Likelihoods for Shifts in Bernoulli	40
3.5.1	Review of Bernoulli Generalized Likelihood Statistic	40
3.5.2	Bernoulli Mixture Likelihood Statistic	41

3.5.3	Consecutive Procedure	42
3.5.4	Fixed Window Procedure	45
3.6	One Edge Increase in General Ising Models and Implications	48
3.6.1	Problem Setting	48
3.6.2	Extended Algorithm from Subsection 3.2.2 and its Performance	49
3.6.3	Implications	50
3.7	Some Lemmas	52
Chapter 4	QCD for Edge Decreased in Ising Models on a Forest	53
4.1	One Edge Decreased/One Component Increased with Known Pre-change Distribution	54
4.1.1	Problem Setting	54
4.1.2	Proposed Algorithm	54
4.1.3	Performance of Algorithm	55
4.2	One Edge Decreased/One Component Increased with only Sign of Pre-change Distribution	57
4.2.1	Problem Setting	57
4.2.2	Proposed Algorithm	58
4.2.3	Performance of Algorithm	59
4.3	Representative Procedure for One Edge Decreased/One Component Increased	62
4.3.1	Proposed Procedure	62
4.3.2	Performance of Algorithm	63
4.3.3	Choosing the Representatives	65

4.4	One Edge Decreased in General Ising Models and Implications	67
4.4.1	Extension from Section4.1	67
4.4.2	Discussion of Extension from Section4.2	68
4.5	Some Lemmas	69
Chapter 5	Conclusion	71
5.1	Summary	71
5.2	Future Work	72
References		75





List of Figures

2.1	Illustration for Markov Property.	18
3.1	Illustration for one edge increase in Ising models on a forest.	35
3.2	Illustration for absolute-valued correlation matrix.	36





Denotation

T	Quickest change detection procedure
τ	Open-ended sequential hypothesis testing procedure
ν	The true change point time.
$x^{(i)}$	Scalar valued data at time i .
$\mathbf{x}^{(i)}$	Vector valued data at time i .
C	Threshold for stopping rules.
\mathbb{E}_i	The expectation on data sequence conditioned on the fact that the change point is at time i .
P_i	The probability on data sequence conditioned on the fact that the change point is at time i .
f_0	The true pre-change distribution.
f_1	The true post-change distribution.
$D(f_1 f_0)$	Kl-divergence between f_1 and f_0 : $D(f_1 f_0) = \mathbb{E}_{X \sim f_1} [\log \frac{f_1(X)}{f_0(X)}]$

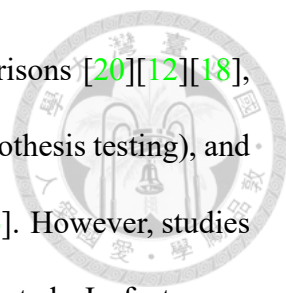




Chapter 1 Introduction

Networks are fundamental structures capable of describing relationships or correlations among a group of entities. Prevalence of networks in real-world applications for instance in communication engineering - sensor networks, financial markets - stock dependency networks, and genomics - gene expression networks have led researchers to study problems such as learning the structure/behavior of the network or making inferences based on knowledge of the underlying network. These problems are particularly important when the network is static. On the other hand, when networks are allowed to change over time, which is usually the case in practice, the timestamp at which the network structure changes or how it changes may be more informative than the network structure itself. For example, in sensor networks, the change in correlation among sensor measurements may imply the eruption of abnormal events; in financial markets, acknowledgement of the change in stock dependency network better adjusts trading strategies; in genomics, identifying the change of gene interactions delineates the biological effects due to external stimuli. Upon mentioned examples motivates the need for rigorous study of structural change detection in a timely manner. In this thesis, the problem will be studied through the lens of a fundamental statistics problem - quickest change detection.

The quickest change detection problem is the study of detecting abrupt statistical changes in streaming data. Classically, researchers laid a strong foundation of the problem

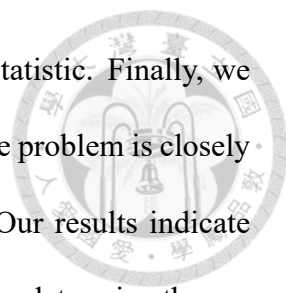


by, for instance, defining appropriate frameworks for future comparisons [20][12][18], establishing connections to related research areas (e.g. sequential hypothesis testing), and answering fundamental questions under the proposed frameworks [16]. However, studies in that era focused only on the case where one data stream is presented. In fact, even until recently, to the best of the author's knowledge, when the quickest change detection problem is studied with the presence of multiple data streams, data streams are either assumed to be mutually independent or almost mutually independent (e.g. only one pair of nodes aren't mutually independent), leaving out challenges and opportunities posed by the existing spatial dependencies. (Note that emphasis of recent studies is varied; e.g. works emphasizing scalability [15][11]; works with compromised observations [6][8].)

In this thesis, to properly incorporate structural relationships for data streams collected over a network, observations are modeled by undirected graphical models [10]; undirected graphical models are a rich class of joint distributions capturing conditional dependencies among random variables based on the notion of reachability on graphs. The strong representation power of such models renders its widespread usage in real-world applications despite its innate computational difficulties for statistical inferences [25]. Under the context of statistical hypothesis testing or parameter estimation (structure learning), studies are usually targeted towards the specific instances of undirected graphical models; Gaussian graphical models for continuous data and Ising models for discrete data. The latter is the subject of this thesis. As a subset of undirected graphical models, inference problems on general Ising models are non-trivial to solve exactly due to the notorious partition function. Researches from the past have suggested that these difficulties arise from the presence of long-range correlations [2], prompting us to first limit ourselves to the tractable graphical models on forests.

Our first attempt at studying the structural changing Ising model quickest change detection problem restricted to distributions on forests has led us to essentially study the correlation change detection problem for disjointed components. We emphasize that intrinsic to the problem, two major challenges are presented; the location at which the change of correlation happens and the difference in correlation before and after the change are both unknown. Specifically,

1. In chapter 3, we study the quickest change detection problem with both the pre-change and post-change distributions belonging to a Ising model on a forest but differing with multiple edges more after the change. Equivalently, the problem is the same as detecting a decrease in the number of tree components in the forest. Our results reveal that with only the pre-change structure information, the problem can be transformed into the pre-change known, post-change unknown Bernoulli quickest change detection problem. Following the well-known procedure from composite quickest change detection, generalized CuSum, we derived a procedure that locates the changes by doing brute force search and distinguishes whether or not a change is presented by using the generalized likelihood ratio for Bernoulli random variables. The procedure is shown to be asymptotically first-order optimal after slight modification of Lorden's result [12]. We further, by taking advantage of the correlation propagation characteristic of the model, show that monitoring only a few selected edges instead of every edge in the entire forest results in a trade-off between computational complexity and statistical guarantees. A rigorous study of the trade-off hints at how we should judiciously choose the representative edges to monitor so as to lower the complexity but maintain decent statistical performance. We then follow the same theme of trading statistical guarantee for computational benefits and



give alternatives to the Bernoulli generalized likelihood ratio statistic. Finally, we extend the study to general Ising models and show that again the problem is closely related to the Bernoulli quickest change detection problem. Our results indicate that the loopy behaviors among the network aggregate to locally determine the parameters for the equivalent Bernoulli quickest change detection problem, thereby explaining why instance optimality can be achieved previously with only structural information. As a corollary, we also show first-order asymptotic instance optimality with only structure information for detection of one edge increase for disconnected components (not necessary trees).

2. In chapter4, we study the dual of the problem from the previous chapter; that is, the problem where the post-change distribution still belonging to a Ising model on a forest differs from the pre-change by having one edge less. This problem is equivalent to the quickest change detection problem for an increase of the number of tree components in the forest by one. We show that when the pre-change distribution is known, the problem is essentially a quickest change detection problem with multiple post-change distributions. Furthermore, we show that generalized CuSum is asymptotically first-order optimal. When the weights of the pre-change distribution are unknown, but the sign and support are known, the problem can be transformed into the pre-change unknown, post-change known Bernoulli quickest change detection problem. By using results from Mei [14] and monitoring every single edge presented in the forest, we offer a procedure with detection delay a constant multiple from optimal. Similar to the previous chapter, correlation propagation over the network provides the opportunity of detecting only a few selected edges with the cost of moderate statistical performance. We quantify this loss and suggest how one should

choose the representative set of edges to monitor. Finally, our extended study for general Ising models reveals its close relationship to the Bernoulli quickest change detection problem. Consequently, we show that our modified Mei's algorithm can be extended to a class of distribution broader than Ising models on a forest.

1.1 Contribution of Thesis

Incorporating complex data distributions and detecting complicated relationship change is a recent theme for the quickest change detection problem [27]. In this thesis, unlike most network quickest change detection problems neglecting spatial dependencies, we focus on quantifying the trade-off between detection delay and false alarm when Markov random fields (probability distributions with correlations described by their structure) undergo a structural change. Structure change in Markov random fields coincides with the change in conditional dependency and has been discussed in offline change point detection [1] and differential network estimation [5]. Our work takes the structural changing assumption to the online setting yielding two opportunities from the perspective of quickest change detection problems.

In chapter 3, we show the first opportunity: the pre-change structure information is sufficient to detect structural changes in a forest network up to asymptotic first-order instance optimality. In contrast, for general composite quickest change detection problems, more precise knowledge of the distributions is often required. The fundamental reason explaining the opportunity lies in the fact that in the likelihood ratios, common structures among the pre/post-change distributions cancel out entirely, leaving only terms involved with the change. Due to the forest structure, the remaining terms in the likelihood ratios are

independent and each behaves like simple one-dimensional random variables after appropriate transformation. We conclude by using known asymptotic first-order optimal results from one dimensional quickest change detection problems and arguing that the presence of multiple independent terms only inflates higher-order terms but leaves first-order terms intact.

Compared with the result from chapter 3, an interesting question to ask: Why in chapter 4, pre-change structural information is no longer sufficient for an asymptotically first-order optimal procedure? A critical reason is the fact that the kl-divergence from chapter 3 depends both on the pre-change weights and the location where the change happens, while the kl-divergence from chapter 4 depends only on the post-change parameters. As conveyed in [14], this complicates the problem and may require a more conservative definition of optimality. In the same work, a more conservative definition was given and procedures achieving such optimality are proposed. Our algorithm from chapter 4 differs from their procedure but still achieves optimal performance in the more conservative sense of Mei.

The second opportunity: paying the price of inferior statistical performance allows for a much more efficient detection scheme, is demonstrated in both chapters 3 and 4. At a high level, the efficient detection scheme gives up from locating precisely the location of change but instead monitors only a pre-chosen set of locations. By the correlation propagation nature of the network, the change will eventually be detected even when the location of the change is still not precisely known. To get the best out of the correlation decay property, we formulate the node selection problem as a maximin optimization problem. The solution to the optimization problem is a node selection algorithm based on finding the gravity center of the pre-change tree structure. It turns out the same algorithm

can be applied in both chapter 3 and chapter 4.







Chapter 2 Backgrounds

In this chapter, two distinct research topics - quickest change detection and undirected graphical models, will be reviewed. The quickest change detection problem is the study of detecting a change in distribution with sequential observations. Undirected graphical models are a class of joint distributions which captures conditional dependency via a graph structure. Here we give a one sentence explanation of why introducing two such distinct topics: The quickest change detection framework allows for rigorous study of change for structured distributions which are naturally modelled by undirected graphical models.

2.1 Quickest Change Detection (QCD)

The quickest change detection problem consists of three entities: a stochastic observation sequence $\{x^{(i)}\}_{i \in \mathcal{N}}$, an unknown change point ν at which statistical properties undergo a change, and a decision-maker who based on its stopping rule T and sequentially collected observations, tries to detect the change as soon as possible. Specifically, the stochastic observation sequence before the change point is assumed to be i.i.d. sampled from the pre-change distribution P_0 (denote using f_0 its mass function or density) and observations after the change point are sampled i.i.d. from the post-change distribution P_1 (denote using f_1 its mass function or density). In the most basic setting, both pre/post-

change distributions are assumed to be known by the decision-maker. At each time-stamp among collecting a new observation, the decision-maker, following its stopping rule, decides based only on past observations whether to stop and announce a change or to keep sampling new observations. Upon stopping, the decision-maker either stops after the true change point $T > \nu$ or triggers a false alarm $T < \nu$. In order to qualify as a "good" decision-maker, the decision-maker should stop with a small detection delay while having false triggers controlled. Lorden's minimax framework [12] introduces two quantities to rigorously evaluate the performance of the decision-maker.

Since the change point is deterministic but unknown, the worst average detection delay of a decision-maker is evaluated by considering the worst possible change point and past realizations.

Definition 2.1.1 (Worst Average Detection Delay).

$$WADD(T) = \sup_{\nu \geq 1} \text{ess sup } \mathbb{E}_\nu[(T - \nu)^+ | x^{(1)}, x^{(2)}, \dots, x^{(\nu-1)}]. \quad (2.1)$$

The average run length to false alarm of a decision-maker is evaluated by the mean of the stopping rule under the pre-change distribution. Controlling the mean is equivalent to controlling the cumulative distribution function (cdf), which under this context, is the probability of false alarm.

Definition 2.1.2 (Average Run Length to False Alarm).

$$ARL2FA(T) = \mathbb{E}_\infty[T]. \quad (2.2)$$

A good performing stopping rule should have a low worst average detection delay

and a decent (lower bounded) average run length to false alarm.



Definition 2.1.3 (Lorden's Problem).

$$\min_{\{T: \text{ARL2FL}(T) \geq \alpha\}} \mathbf{WADD}(T). \quad (2.3)$$

Lorden's optimization problem is solved asymptotically, as $\alpha \rightarrow \infty$. Specifically, he characterized the fundamental limit of the optimization problem - the converse result, and showed that an efficient algorithm - CuSum achieves the limit.

Theorem 2.1.1 (Converse of Lorden's Problem). *Every stopping rule T satisfies $\mathbf{WADD}(T) \geq (1 + o(1)) \frac{\log \text{ARL2FA}(T)}{D(f_1 \| f_0)}$.*

Corollary 2.1.1. *Any stopping rule T satisfying $\text{ARL2FL}(T) \geq \alpha$ must have $\mathbf{WADD}(T) \geq \frac{\log \alpha(1+o(1))}{D(f_1 \| f_0)}$ as $\alpha \rightarrow \infty$.*

Theorem 2.1.2 (Achievability of Lorden's Problem). *Page's CuSum algorithm T_{CuSum} [17] is asymptotically optimal. (i.e. $\text{ARL2FA}(T_{\text{CuSum}}) \geq \alpha$ and $\mathbf{WADD}(T_{\text{CuSum}}) \leq \frac{\log \alpha(1+o(1))}{D(f_1 \| f_0)}$ as $\alpha \rightarrow \infty$.)*

The CuSum algorithm detects the change by locating a starting time-stamp where cumulative sums of log-likelihood ratios admit a strong positive drift. CuSum algorithm reads:

$$T(C_\alpha = \log \alpha) = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^n \log \frac{f_1(x^{(i)})}{f_0(x^{(i)})} \geq C_\alpha \right\}. \quad (2.4)$$

Intuitively, the algorithm works since under pre/post-change distribution, the sum of log-likelihood ratios has a positive/negative drift.

In practice, the CuSum algorithm is implemented using its recursive form:

$$W_1 = 0; W_{n+1} = \max(0, W_n + \log \frac{f_1(x^{(n+1)})}{f_0(x^{(n+1)})}). \quad (2.5)$$

$$T(C_\alpha) = \inf\{n \geq 1 : W_n \geq C_\alpha\}.$$



Before proceeding, here's a sketch of how Lorden proved his achievability result by relating results from the open-ended sequential probability ratio test from the context of sequential hypothesis testing.

Theorem 2.1.3 (Lorden's Theorem). *Let τ be an **extended stopping time (open-ended stopping time)** such that $P_\infty(\tau < \infty) \leq \alpha$ and let τ_k be the stopping time obtained by applying τ to $x^{(k)}, x^{(k+1)}, \dots$. Define $T = \min\{\tau_k + k - 1 : k = 1, 2, \dots\}$. Then*

$$ARL2FA(T) \geq \frac{1}{\alpha},$$

and for any alternative distribution F_1 ,

$$WADD(T) \leq \mathbb{E}_0[\tau].$$

Briefly speaking, an open-ended stopping time in the context of sequential hypothesis testing is a stopping rule which stops only when the underlying observation sequence looks like the alternative hypothesis; when the underlying sequence is in fact the null hypothesis, it may not stop. According to Lorden's theorem, to prove the performance of CuSum algorithm, it suffices to analyze the performance of its associated open-ended stopping time - open-ended sequential probability ratio test.

Definition 2.1.4 (Open-ended Sequential Probability Ratio Test).

$$\tau(C) = \inf \left\{ n \geq 1 : \sum_{i=1}^n \log \frac{f_1(x^{(i)})}{f_0(x^{(i)})} \geq C \right\}. \quad (2.6)$$



Lemma 2.1.1 (Upper bound to $P_\infty(\tau \leq \infty)$).

$$P_\infty(\tau < \infty) \leq 2^{-C}. \quad (2.7)$$

Lemma 2.1.2 (Lower bound to $\mathbb{E}_0(\tau)$).

$$\mathbb{E}_0[\tau] \geq \frac{C(1 + o(1))}{D(f_1||f_0)} \text{ as } C \rightarrow \infty. \quad (2.8)$$

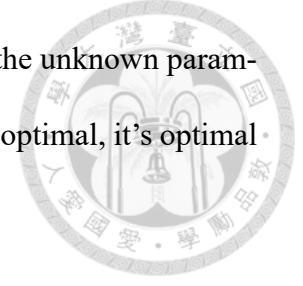
Lemma 2.1.1 is application of maximal inequality for martingales and lemma 2.1.2 derives from Wald's identity. Both results can be found in the nice tutorial [24].

Theorem 2.1.3, lemma 2.1.1, and lemma 2.1.2 concludes the proof of achievability for the CuSum algorithm. It's worthwhile to notice that based on Lorden's result, analyzing the performance of quickest change detection procedures is not much different from analyzing that of open-ended sequential hypothesis testing procedures.

2.1.1 Composite QCD

Previously, the quickest change detection problem is studied under the scenario that both the pre/post-change distributions are specified by the decision-maker. In this section, the rather practical problem where unknowns are presented in the pre/post-change distributions will be addressed. We further emphasize that when the problem is composite, two types of optimality occurs - **instance optimal** and **robust optimal**. When a procedure is

instance optimal, it means that the procedure performs as good as if the unknown parameter is known to the decision-maker, whereas if a procedure is robust optimal, it's optimal only in a minimax sense.



2.1.1.1 Post-change Unknown QCD

When the distribution after the change is not exactly known to the decision-maker, the problem belongs to the post-change unknown QCD problem. Such extension to the original QCD problem is probably the most common in the literature [22][26]. Although algorithms of different forms are proposed in the literature, most of which are variants to the generalized or mixture approach.

Evaluation of the performance under the current setting inherits from Lorden's original problem 2.1.3. A subtle difference lies in the fact that the essential supremum taken in $\mathbf{WADD}(\cdot)$ now depends on the underlying post-change distribution while the decision-maker T should not.

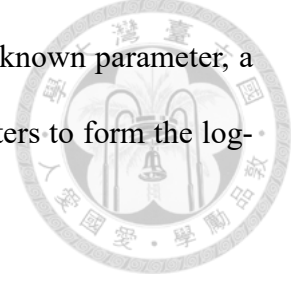
The generalized approach deals with the unknown by estimating the unknown parameter and subsequently plugging in the estimator to form the log-likelihood ratio.

Definition 2.1.5 (Generalized CuSum).

$$T_G = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{\theta \in \Theta} \left[\sum_{i=k}^n \log \frac{f_{\theta}(x^{(i)})}{f_{\theta_0}(x^{(i)})} \right] \geq C \right\}. \quad (2.9)$$

Using this procedure, Lorden [12] showed that for a shift of the parameter in one-parameter exponential families, the test is asymptotically **instance optimal**. Later, Siegmund [21] presents a detailed approximation for the problem of Gaussian mean shift.

For the mixture approach, instead of striving to estimate the unknown parameter, a non-degenerate weight is placed on all possible post-change parameters to form the log-likelihood ratio.



Definition 2.1.6 (Mixture CuSum).

$$T_M = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \log \int \prod_{i=k}^n \frac{f_\theta(x^{(i)})}{f_0(x^{(i)})} dG(\theta) > C \right\}. \quad (2.10)$$

Using Lorden's theorem and Pollak's result [19], the test is asymptotically **instance optimal** for one-parameter exponential family parameter shift.

Since when a procedure is asymptotically instance optimal, it performs as good as if the unknown parameter is known to the decision-maker asymptotically. One might ask, then what is the price with the existence of the unknowns? Statistically, both procedures have strictly larger higher-order detection delays ($o(1)$ terms in $\frac{\log \alpha(1+o(1))}{D(f_1||f_0)}$); computationally, both procedures are no longer recursive in general.

2.1.1.2 Pre-change Unknown QCD

The problem of detecting a change from a set of potential pre-change distributions to a specific, known post-change distribution is less studied in the literature. Mei [14] showed for one-sided shift of the the parameter in one-parameter exponential distributions is asymptotically **reverse instance optimal**. The term **reverse** means the role of **ARL2FA** and **WADD** is reversed.

Definition 2.1.7 (Reverse Lorden's Problem).

$$\max_{\{T: \mathbf{WADD}(T) \leq \alpha\}} \mathbf{ARL2FA}(T).$$

Definition 2.1.8 (Mei's procedure).

$$T_{Mei}(a) = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^n \log \frac{f_1(x^{(i)})}{f_\delta(x^{(i)})} \geq D(f_1 || f_\delta) a \text{ for all } \infty < \delta_- \leq \delta \leq \delta_+ < \delta_1 \right\}. \quad (2.11)$$

The pre-change distribution parameter space is disjoint of the true post-change parameter δ_1 .

Under the pre-change unknown setting, Mei's procedure exhibits the same asymptotically instance optimal property as in the post-change unknown setting (i.e. both achieves Lorden's converse); however, the two settings are fundamentally different in the sense that the role of **ARL2FA** and **WADD** is reversed.

2.1.1.3 Pre/Post-change Unknown QCD

When both the pre/post-change distributions are not completely specified, calling for instance optimality can be too strict in general. Taking one step back, [23] tackled the problem under a minimax problem setting and derived **robust optimal** procedures based on identifying the least favorable distribution in the composite pre/post-change parameter set.

Definition 2.1.9 (Robust optimal criteria).

$$\min_{\{T: \min_{f_0 \in \mathcal{F}_0} \text{ARL2FL}(T) \geq \alpha\}} \max_{f_0 \in \mathcal{F}_0; f_1 \in \mathcal{F}_1} \text{WADD}(T). \quad (2.12)$$

Another less conservative definition of optimality is given by Mei [14].

Definition 2.1.10 (Asymptotic efficient). *A stopping rule T is asymptotic efficient at pre/*

post-change distribution pair (f_0, f_1) if under (f_0, f_1) ,

$$\liminf_{C \rightarrow \infty} \frac{\log \mathbf{ARL2FA}(T)}{D(f_1 || f_0) \mathbf{WADD}(T)} = 1.$$



Definition 2.1.11 (Mei's optimal criteria). *A stopping rule T is asymptotically optimal to first-order if:*

1. *for each $f_0 \in \mathcal{F}_0$, there exists at least one $f_1 \in \mathcal{F}_1$ such that T is asymptotically efficient at $(f_0, f_1(f_0))$; and*
2. *for each $f_1 \in \mathcal{F}_1$, there exists at least one $f_0 \in \mathcal{F}_0$ such that T is asymptotically efficient at $(f_0(f_1), f_1)$.*

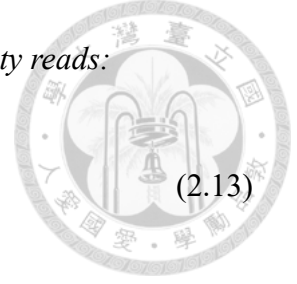
2.2 Undirected Graphical Models

Modeling a large set of random variables is no easy task; it's more so if the notion of causality or correlation is not taken into consideration. Appropriate modeling of a set of random variables using causality gives rise to a class of joint distributions called directed graphical models (also called Bayesian networks) in the literature. Its counterpart class of distributions, undirected graphical models (also called Markov random fields or Gibbs random field), which are modeled based on correlation, is the topic of this section. A specific subset of undirected graphical models, Ising models, will be used throughout this thesis to model network behavior.

Quite often, the joint behavior of a group of variables inherits from local interactions within subsets of variables. This is exactly how Gibbs random fields are defined.

Definition 2.2.1 (Gibbs Random Fields). *The mass function or density reads:*

$$f(\mathbf{x}) = \frac{1}{Z} \prod_{\text{clique } C} \psi_C(\mathbf{x}_C) \quad (2.13)$$



The $\psi_C(\cdot)$ are potential functions that give a non-negative score for each maximal clique. The scores quantify local interactions among variables within the clique. Aggregation of local interactions to form a global score is done by doing multiplication over all maximal cliques. Finally, the partition function Z serves as the normalization constant so as to turn $f(\cdot)$ into a valid distribution.

It turns out distributions with such factorization models conditional dependency via reachability on graphs. (Formally, this is the Hammersley-Clifford theorem.)

Definition 2.2.2 (Markov Random Fields). *Markov random fields are joint distributions which satisfy markov properties.*

To demonstrate the Markov property, refer to figure 2.1. The black variable is conditionally independent of the grey ones given the white ones since the black node is no longer connected to the grey nodes after removing the white nodes.

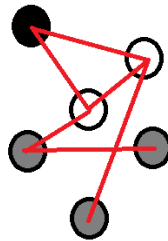


Figure 2.1: Illustration for Markov Property.

Despite being equipped with powerful representation power, inference problems such as computation of marginals/modes and parameter estimation on undirected graphical

models are known to be non-trivial [25], mainly due to the notorious partition function.



2.2.1 Ising Models

In the field of statistical hypothesis testing or parameter estimation (structure learning), the complete family of undirected graphical models may be too general to study. Quite often in the literature, the study is restricted to either the continuous gaussian graphical model or the discrete Ising model. An introduction to Ising models will be presented in the following.

Ising model is a discrete (each random variable takes value $+1$ or -1), parametric joint distribution which models pairwise interaction.

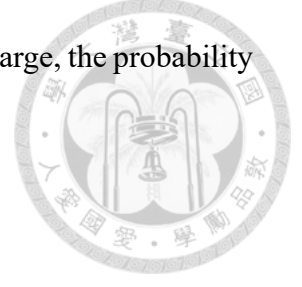
Definition 2.2.3 (Ising Model). *Joint probability mass function reads:*

$$f(\mathbf{x}) = \frac{1}{Z} \exp\left\{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t\right\}. \quad (2.14)$$

The partition function normalizes the probability $Z = \sum_{x_1, x_2, \dots, x_d = \{-1, +1\}^d} \exp\{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t\}$. Note that since the Ising model is a particular instance of an undirected graphical model, it exhibits the Markov property (through the edge set \mathcal{E}).

The model which originates from statistical physics to study the behavior of phase transitions was later widely discussed in the scientific community. Its parity nature is perfect for the modeling of a wide variety of systems including spin up/down in physics, neuron activated/deactivated in neural science, or upvotes/downvotes in social networks. The presence of the mean-field vector δ introduces the bias of each individual. (i.e. when δ_s is large, the probability of $x_s = +1$ is large and vice versa.) The interaction parameters

θ_{ut} govern how neighboring nodes affect each other. (i.e. when θ_{ut} is large, the probability of x_u and x_t taking the same sign is large, and vice versa.)



2.2.1.1 Ising Model in a Forest

When the set of variables in Ising model exhibit a natural hierarchical structure, the underlying edge set \mathcal{E} can be conveniently assumed to be a tree graph. Furthermore, if multiple disconnected hierarchical components are presented, then, jointly, the distribution belongs to Ising model in a forest. By removing cycles from the edge set \mathcal{E} and assuming zero mean-field vector $\delta = 0$, the model admits a simple probability mass function.

Definition 2.2.4 (Ising Model in a Forest). *Joint probability mass function reads:*

$$f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j). \quad (2.15)$$

The two restrictions sacrifice representation power for tractable properties of likelihood functions and correlations.

Fact 2.2.1 (Equal Marginal Probabilities).

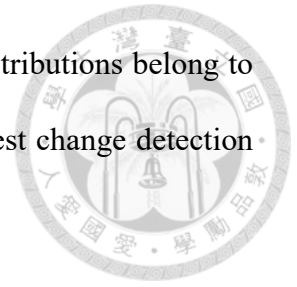
$$\Pr(x_u = +1) = \Pr(x_u = -1) = \frac{1}{2} \quad \forall u \in \mathcal{V}. \quad (2.16)$$

Fact 2.2.2 (Correlation Decay Property).

$$\mathbb{E}[x_u x_v] = \prod_{\mu \in \text{path}(u,v)} \tanh(\theta_\mu). \quad (2.17)$$

Simple as it is, there have been numerous study of the model for structure learning [9][3] and hypothesis testing [7][4]. In the following chapters, the quickest change de-

tection problem will be examined first under the assumption that distributions belong to Ising model in a forest. This should provide intuition for the quickest change detection problem under the general Ising model.







Chapter 3 QCD for Edge Increased in Ising Models on a Forest

In this chapter, we study the first type of structural change in Ising models on a forest: edge appearance. Our expedition starts off with the simplest case where the decision-maker knows the pre-change weights over the forest and only one edge pops out after the change happens. Results quickly reveal that knowledge of only the pre-change structure is sufficient for achieving asymptotic first-order optimality. We further show that our algorithm, based on generalized CuSum, can be extended to the setting where multiple edges appear after the change. The performance of the algorithm remains asymptotically first-order optimal albeit inflation of the higher-order terms. Next, we turn our attention to computational considerations. We show that instead of monitoring every possible edge increase, as in generalized CuSum, we can take advantage of the correlation propagation characteristic of the model and further provide a more efficient edge sampling scheme at the expense of some statistical performance. Furthermore, we give computationally simpler but statistically inferior alternatives of the generalized likelihood ratio statistic, which can subsequently be applied to edges chosen to be monitored. Finally, in the last section, we show results for one edge increase in general models and discuss its implications.



3.1 One Edge Increased/One Component Decreased with Known Pre-Change Distribution

3.1.1 Problem Setting

The decision-maker T collects data among the network in a timely manner. The decision-maker does not know when the underlying data sequence undergoes a change. The goal would be to raise an alarm as soon as the decision-maker believes the change has happened. Here we assume that:

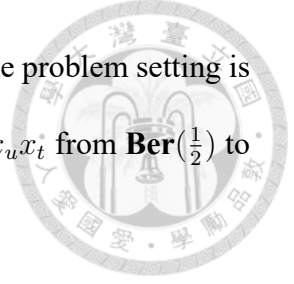
1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from **known** distribution f_0 belonging to Ising model on forest (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j)$).
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution f_1 differing from pre-change distribution by one edge **more**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}^+} (1 + \tanh(\theta_{ij})x_i x_j)$).

Note that before the change, edge set \mathcal{E} and parameters $\theta_{ij} \forall \{i, j\} \in \mathcal{E}$ are completely specified. After the change, edge set $\mathcal{E}^+ = \{\mathcal{E}, \{r, s\}\}$ for some $\{r, s\} \in \mathcal{E}^{C;\text{forest}}$; parameters $\theta_{ij} \in \mathcal{E}$ remains untouched, but θ_{rs} is unknown.

3.1.2 Proposed Algorithm

The problem belongs to the composite post-change quickest change detection problem, prompting us to try out generalized CuSum. It turns out, as hinted by likelihood

ratio $\frac{f_1(x)}{f_0(x)} = \frac{1+\tanh \theta_{ut} x_u x_t}{1} = \frac{1+\tanh \theta_{ut} x_u x_t}{\frac{1}{2}}$, generalized CuSum under the problem setting is slight modification to generalized CuSum for detecting a change of $x_u x_t$ from $\mathbf{Ber}(\frac{1}{2})$ to unknown $\mathbf{Ber}(\frac{1+\tanh \theta_{ut} x_u x_t}{2})$.



$$T_G = \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}^C; \text{forest}} \max_{\theta_{rs} \geq |\theta_{\text{threshold}}|} \sum_{i=k}^n \log(1 + \tanh(\theta_{rs}) x_r^{(i)} x_s^{(i)}) \geq C \right\}. \quad (3.1)$$

Here in generalized CuSum, $\theta_{\text{threshold}}$ is a function of the threshold C . We will be choosing $\theta_{\text{threshold}} = \frac{1}{C}$, although other choices may do. Solving inner maximization problem (without constraint) by fixing k and θ_{rs} yields (see proof3.7):

$$\max_{\theta_{rs}} \prod_{i=k}^n (1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}) \Leftrightarrow \theta_{rs}^* = \frac{1}{2} \log\left(\frac{\text{count}_{+,rs}(k, n)}{\text{count}_{-,rs}(k, n)}\right). \quad (3.2)$$

Here $H(\cdot)$ denotes the binary entropy function and $\text{count}_{+,rs}(k, n)/\text{count}_{-,rs}(k, n)$ denotes the number of $+1/-1$ in $\mathbf{x}_r^{(k)} \mathbf{x}_s^{(k)}, \mathbf{x}_r^{(k+1)} \mathbf{x}_s^{(k+1)}, \dots, \mathbf{x}_r^{(n)} \mathbf{x}_s^{(n)}$.

Therefore, if $|\frac{1}{2} \log(\frac{\text{count}_{+,rs}(k, n)}{\text{count}_{-,rs}(k, n)})| \geq |\theta_{\text{threshold}}|$, maximum attains at $\frac{1}{2} \log(\frac{\text{count}_{+,rs}(k, n)}{\text{count}_{-,rs}(k, n)})$ and the corresponding score is $(n - k + 1)(1 - H(\frac{\text{count}_{+,rs}(k, n)}{n - k + 1}))$; otherwise, maximum attains at either $\theta_{\text{threshold}}$ or $-\theta_{\text{threshold}}$ and the corresponding score is $\theta_{\text{threshold}}(\text{count}_{+,rs}(n, k) - \text{count}_{-,rs}(n, k)) - (n - k + 1) \log(\cosh \theta_{\text{threshold}})$.

The more cautious reader might have noticed that the procedure T_G based on generalized CuSum does not depend on the pre-change weights. This suggests that good performance may be achieved with only structure information from the pre-change distribution. We will quantify this in the next section (section3.2). Results there immediately imply that generalized CuSum T_G is asymptotically first-order optimal under the problem setting of this section.



3.2 One Edge Increased/One Component Decreased with only Structure of Pre-Change Distribution

3.2.1 Problem Setting

Under the quickest change detection problem setting, we further assume that:

1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from distribution f_0 belonging to Ising model on forest (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j)$). **Parameters θ_{ij} are not known but support \mathcal{E} is known.**
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution f_1 differing from pre-change distribution by one edge **more**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}^+} (1 + \tanh(\theta_{ij})x_i x_j)$).

3.2.2 Proposed Algorithm

We use the same procedure from the previous section.

$$T_G = \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}^C; \text{forest}} \max_{\theta_{rs} \geq |\theta_{\text{threshold}}|} \sum_{i=k}^n \log(1 + \tanh(\theta_{rs})x_r^{(i)}x_s^{(i)}) \geq C \right\}. \quad (3.3)$$

We choose $\theta_{\text{threshold}} = \frac{1}{C}$, although other choices may do. Solution to the inner maximization problem says if $|\frac{1}{2} \log(\frac{\text{count}_{+;rs}(k,n)}{\text{count}_{-;rs}(k,n)})| \geq |\theta_{\text{threshold}}|$, maximum attains at $\frac{1}{2} \log(\frac{\text{count}_{+;rs}(k,n)}{\text{count}_{-;rs}(k,n)})$ and the corresponding score is $(n - k + 1)(1 - H(\frac{\text{count}_{+;rs}(n,k)}{n-k+1}))$; otherwise, maximum attains at either $\theta_{\text{threshold}}$ or $-\theta_{\text{threshold}}$ and the corresponding score is $\theta_{\text{threshold}}(\text{count}_{+;rs}(n, k) - \text{count}_{-;rs}(n, k)) - (n - k + 1) \log(\cosh \theta_{\text{threshold}})$. For details

see subsection 3.1.2.



3.2.3 Important Property

Before showing the performance of the algorithm, we present an important property of the generalized CuSum maximization problem [12]. The property shows that the mapping from the stopped time to the parameter attaining the maximum is bijective (when restricting the parameter space to either positive or negative real line), easing the proof for the false alarm. Specifically, we consider the inner maximization problem:

$$\max_{\theta \geq |\theta_{threshold}|} \log((1 + \tanh \theta)^{\text{count}_+} (1 - \tanh \theta)^{\text{count}_-}) \quad (3.4)$$

$$= \max_{\theta \geq |\theta_{threshold}|} \log \exp\{\theta(\text{count}_+ - \text{count}_-) - n \log(\cosh \theta)\}. \quad (3.5)$$

The procedure stops only when $\max_{\theta > |\theta_{threshold}|} \theta(\text{count}_+ - \text{count}_-) - n \log(\cosh \theta) \geq C$,
or

$$\begin{cases} \text{count}_+ - \text{count}_- \geq \min_{\theta \geq \theta_{threshold}} \left\{ \frac{C}{\theta} + \frac{n \log(\cosh \theta)}{\theta} \right\}. \\ \text{count}_+ - \text{count}_- \leq \max_{\theta \leq -\theta_{threshold}} \left\{ \frac{C}{\theta} + \frac{n \log(\cosh \theta)}{\theta} \right\}. \end{cases}$$

Define $h(\theta) = \frac{C}{\theta} + \frac{n \log(\cosh \theta)}{\theta}$, then $h'(\theta) = -\frac{C}{\theta^2} + \frac{n \tanh \theta}{\theta} - \frac{n \log(\cosh \theta)}{\theta^2}$. Minimum/maximum attained at $h'(\theta) = 0$, yielding $\theta^* = g^{-1}(\frac{C}{n})$ (where $g(\theta) = \theta \tanh \theta - \log(\cosh \theta) = 1 - H(\frac{1+\tanh \theta}{2})$) if $n \leq \frac{C}{1-H(\frac{1+\tanh \theta_{threshold}}{2})}$. The important implication is if the procedure stops at time $n \leq \frac{C}{1-H(\frac{1+\tanh \theta_{threshold}}{2})}$, the parameter attained must be $g^{-1}(\frac{C}{n})$ or $-g^{-1}(\frac{C}{n})$; if the procedure stops at time $n > \frac{C}{1-H(\frac{1+\tanh \theta_{threshold}}{2})}$, the parameter attained is $\theta_{threshold}$.

3.2.4 Performance of Proposed Algorithm



Theorem 3.2.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$\mathbf{WADD}(T_G) \leq \frac{C(1 + o(1))}{D(f_1||f_0)} \text{ as } C \rightarrow \infty. \quad (3.6)$$

$$\mathbf{ARL2FA}(T_G) \geq (1 + o(1))2^C. \quad (3.7)$$

Proof. To facilitate the proofs, define $\tau_G = \inf \{n \geq 1 : \max_{(r,s) \in \mathcal{E}^C; \text{forest}} \max_{\theta_{rs} \geq |\theta_{\text{threshold}}|} \sum_{i=1}^n \log(1 + \tanh(\theta_{rs})x_r^{(i)}x_s^{(i)}) \geq C\}$ SPRT using the true post-change parameter $\tau_{rs} = \inf \{n \geq 1 : \sum_{i=1}^n \log(1 + \tanh \theta_{rs}x_r^{(i)}x_s^{(i)}) \geq C\}$, and generalized SPRT using location (r, s) $\tau_{G;rs} = \inf \{n \geq 1 : \max_{\theta_{rs} \geq |\theta_{\text{threshold}}|} \sum_{i=1}^n \log(1 + \tanh \theta_{rs}x_r^{(i)}x_s^{(i)}) \geq C\}$.

Proof for WADD:

By Lorden's theorem 2.1.3, $\mathbf{WADD}(T_G) \leq \mathbb{E}_0[\tau_G]$. Furthermore,

$$\mathbb{E}_0[\tau_G] \leq \mathbb{E}_0[\tau_{rs}] = \frac{C(1 + o(1))}{D(f_1||f_0)} \text{ as } C \rightarrow \infty.$$

First inequality follows from the fact that eventually $|\theta_{\text{threshold}}| < |\theta_{rs}|$, so τ_G eventually stops earlier than τ_{rs} . Final inequality derives from Wald's identity.

Proof for ARL2FA:

By Lorden's theorem 2.1.3, to prove $\mathbf{ARL2FA}(T_G) \geq (1 + o(1))2^C$, it suffice to

show $P_\infty(\tau_G < \infty) \leq \frac{(1+o(1))}{2^C}$.



$$\begin{aligned}
 P_\infty(\tau_G < \infty) &= P_\infty(\cup_{(r,s) \in \mathcal{E}^{C;\text{forest}}} \{\tau_{rs} < \infty\}) = |\mathcal{E}^{C;\text{forest}}| P_\infty(\tau_{rs} < \infty) \quad (3.8) \\
 &= |\mathcal{E}^{C;\text{forest}}| \left(\sum_{n=1}^{\frac{C}{1-H(\frac{1+\tanh \theta_{\text{threshold}}}{2})}} P_\infty(\tau_{rs} = n) + P_\infty\left(\frac{C}{1-H(\frac{1+\tanh \theta_{\text{threshold}}}{2})} < \tau_{rs} < \infty\right) \right) \quad (3.9)
 \end{aligned}$$

$$\leq |\mathcal{E}^{C;\text{forest}}| \left(\frac{C}{1-H(\frac{1+\tanh \theta_{\text{threshold}}}{2})} 2^{-C} + 2^{-C} \right) \leq \left(\frac{C}{1-\text{sech } \theta_{\text{threshold}}} 2^{-C} + 2^{-C} \right) \quad (3.10)$$

$$\leq |\mathcal{E}^{C;\text{forest}}| \left(\frac{4C}{\theta_{\text{threshold}}^2} 2^{-C} + 2^{-C} \right) = |\mathcal{E}^{C;\text{forest}}| (4C^3 2^{-C} (1 + o(1))) \quad (3.11)$$

First inequality is due to the important bijection between stopped time-stamp and parameter attaining maximum (see 3.2.3). Second inequality derives from upper bound to the binary entropy function $H(p) \leq (4p(1-p))^{\frac{1}{2}}$ and final inequality derives from $\text{sech } \theta \leq 1 - \frac{\theta^2}{4}$ over $\theta < 1$ and our selection $\theta_{\text{threshold}} = \frac{1}{C}$. \square

Theorem 3.2.2. *Taking $C = \log \alpha$, achievability result is established:*

$$\mathbf{ARL2FA}(T_G) \geq \alpha; \quad \mathbf{WADD}(T_G) \leq \frac{\log \alpha (1 + o(1))}{D(f_1 || f_0)} \text{ as } \alpha \rightarrow \infty. \quad (3.12)$$

Therefore, proposed procedure T_G is asymptotically first-order instance optimal.

Proof. The first part of the theorem follows directly from theorem 3.2.1. Optimality follows from comparing the converse result of corollary 2.1.1 and the achievability result. \square

3.3 Multiple Edge Increased/One Component Decreased with only Structure of Pre-Change Distribution



3.3.1 Problem Setting

In this section, we extend the setting to where multiple edges appear after the change. Specifically, we assume:

1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from distribution P_0 belonging to Ising model on forest (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j)$). **Parameters θ_{ij} are not known but support \mathcal{E} is known.**
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution P_1 differing from pre-change distribution by s edges **more**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}^+} (1 + \tanh(\theta_{ij})x_i x_j)$).

We further assume that the decision-maker doesn't know the precise number of edges increase s , but upper bound to number of edges increased is known s_{max} .

3.3.2 Proposed Algorithm

The algorithm again is based on generalized CuSum.



$$T_G = \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{e_1, e_2, \dots, e_{s_{max}} \in \mathcal{E}^C; \text{forest}} \max_{\theta_{e_1}, \theta_{e_2}, \dots, \theta_{e_{s_{max}}} \geq |\theta_{threshold}|} \sum_{i=k}^n \log \left((1 + \tanh(\theta_{e_1})[x^{(i)}x^{(i)}]_{e_1})(1 + \tanh(\theta_{e_2})[x^{(i)}x^{(i)}]_{e_2}) \dots (1 + \tanh(\theta_{e_{s_{max}}})[x^{(i)}x^{(i)}]_{e_{s_{max}}}) \right) \geq C \right\}.$$

Same as in previous subsections, we let $\theta_{threshold} = \frac{1}{C}$.

A critical notice is $\{e_1, e_2, \dots, e_{s_{max}}\}$ encompasses only the set of edges such that after adding them, the entire graph remains a forest graph (no loops). For example, when e_1, e_2 connects component one and two, two and three respectively, no edges in $\{e_3, \dots, e_{s_{max}}\}$ should connect components one and three.

3.3.3 Performance of Proposed Algorithm

Theorem 3.3.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$WADD(T_G) \leq \frac{C(1 + o(1))}{D(f_1||f_0)} \text{ as } C \rightarrow \infty. \quad (3.13)$$

$$ARL2FA(T_G) \geq (1 + o(1))2^C. \quad (3.14)$$

Proof. To facilitate the proofs, define $\tau_G(C) = \{n \geq 1 : \max_{e_1, e_2, \dots, e_{s_{max}} \in \mathcal{E}^C; \text{forest}}$

$\max_{\theta_{e_1}, \theta_{e_2}, \dots, \theta_{e_{s_{max}}} \geq |\theta_{threshold}|} \sum_{i=1}^n \log \left((1 + \tanh(\theta_{e_1})[x^{(i)}x^{(i)}]_{e_1})(1 + \tanh(\theta_{e_2})[x^{(i)}x^{(i)}]_{e_2}) \dots (1 + \tanh(\theta_{e_{s_{max}}})[x^{(i)}x^{(i)}]_{e_{s_{max}}}) \right) \geq C \}$, SPRT using the true post-change parameter $\tau_{e_1^*, e_2^*, \dots, e_s^*}(C) =$

$\inf \{n \geq 1 : \sum_{i=1}^n \log ((1 + \tanh(\theta_{e_1}^*)[x^{(i)}x^{(i)}]_{e_1^*})(1 + \tanh(\theta_{e_2}^*)[x^{(i)}x^{(i)}]_{e_2^*}) \dots (1 + \tanh(\theta_{e_s}^*)$
 $[x^{(i)}x^{(i)}]_{e_s^*})) \geq C\}$, generalized SPRT using locations $e_1, e_2, \dots, e_{s_{max}}$: $\tau_{G;e_1,e_2,\dots,e_{s_{max}}}(C) =$
 $\inf \{n \geq 1 : \max_{\theta_{e_1}, \theta_{e_2}, \dots, \theta_{e_{s_{max}}}} \sum_{i=1}^n \log (1 + \tanh(\theta_{e_1})[x^{(i)}x^{(i)}]_{e_1})(1 + \tanh(\theta_{e_2})$
 $[x^{(i)}x^{(i)}]_{e_2}) \dots (1 + \tanh(\theta_{e_{s_{max}}})[x^{(i)}x^{(i)}]_{e_{s_{max}}})) \geq C\}$, and generalized SPRT locating only
 one edge e_1 $\tau_{G;e_1}(C) = \inf \{n \geq 1 : \max_{\theta_{e_1}} \sum_{i=1}^n \log(1 + \tanh(\theta_{e_1})[x^{(i)}x^{(i)}]_{e_1}) \geq$
 $C\}$.

Proof for WADD:

By Lorden's theorem 2.1.3, $\mathbf{WADD}(T_G) \leq \mathbb{E}_0[\tau_G]$. Furthermore,

$$\mathbb{E}_0[\tau_G] \leq \mathbb{E}_0[\tau_{e_1^*, e_2^*, \dots, e_s^*}] = \frac{C(1 + o(1))}{D(f_1 || f_0)} \text{ as } C \rightarrow \infty.$$

First inequality follows from the fact that eventually $|\theta_{threshold}| < |\theta_{e_1^*}|, |\theta_{e_2^*}|, \dots, |\theta_{e_s^*}|$, so τ_G eventually stops earlier than $\tau_{e_1^*, e_2^*, \dots, e_s^*}$. Final inequality derives from Wald's identity.

Proof for ARL2FA:

By Lorden's theorem 2.1.3, to prove $\mathbf{ARL2FA}(T_G) \geq (1 + o(1))2^C$, it suffice to show $P_\infty(\tau_G < \infty) \leq \frac{(1+o(1))}{2^C}$.

Define the integral $F(x) = \int x^3 2^{-x} dx = 2^{-x} \sum_{i=0}^3 (-1)^{3-i} \frac{3!}{i!(-\log_e 2)^{3-i+1}} x^i$; then

$$\int_0^\infty x^3 2^{-x} dx = -F(0) = \frac{3!}{(\log_e 2)^4}. \text{ For large } C,$$



$$\begin{aligned}
& P_\infty(\tau_G(C) < \infty) = P_\infty(\cup_{e_1, e_2, \dots, e_{s_{max}} \in \mathcal{E}^C; \text{forest}} \{\tau_{G; e_1, e_2, \dots, e_{s_{max}}}(C) < \infty\}) \\
& \leq \text{Const.} \times P_\infty(\tau_{G; e_1, e_2, \dots, e_{s_{max}}}(C) < \infty) \\
& = \text{Const.} \times P_\infty(\cup_{C_1 + C_2 + \dots + C_{s_{max}} = C} \{\tau_{G; e_1}(C_1) < \infty, \tau_{G; e_2}(C_2) < \infty, \dots, \tau_{G; e_{s_{max}}}(C_{s_{max}}) < \infty\}) \\
& \leq \text{Const.} \int \int \dots \int_{C_1 + C_2 + \dots + C_{s_{max}} \geq C} P_\infty(\tau_{e_1}(C_1) < \infty) P_\infty(\tau_{e_2}(C_2) < \infty) \\
& \quad \dots P_\infty(\tau_{e_{s_{max}}}(C_{s_{max}}) < \infty) dC_1 dC_2 \dots dC_{s_{max}} \\
& \leq \text{Const.} \int \int \dots \int_{C_1 + C_2 + \dots + C_{s_{max}} \geq C} 4C_1^3 2^{-C_1} \times 4C_2^3 2^{-C_2} \dots \times 4C_{s_{max}}^3 2^{-C_{s_{max}}} dC_1 dC_2 \dots dC_{s_{max}} \\
& = \text{Const.} \left(\int_0^\infty \int_0^\infty \dots \int_0^\infty C_1^3 2^{-C_1} \times C_2^3 2^{-C_2} \dots \times C_{s_{max}}^3 2^{-C_{s_{max}}} dC_1 dC_2 \dots dC_{s_{max}} \right. \\
& \quad \left. - \int_0^C \int_0^{C-C_{s_{max}}} \dots \int_0^{C-C_{s_{max}}-\dots-C_2} C_1^3 2^{-C_1} \times C_2^3 2^{-C_2} \dots \times C_{s_{max}}^3 2^{-C_{s_{max}}} dC_1 dC_2 \dots dC_{s_{max}} \right) \\
& = \text{Const.} \left((-F(0))^{s_{max}} - \int_0^C \int_0^{C-C_{s_{max}}} \dots \int_0^{C-C_{s_{max}}-\dots-C_3} C_2^3 2^{-C_2} \dots \times C_{s_{max}}^3 2^{-C_{s_{max}}} \right. \\
& \quad \left. (F(C - C_{s_{max}} - \dots - C_2) - F(0)) dC_2 \dots dC_{s_{max}} \right) \\
& = \text{Const.} \left((-F(0))^{s_{max}} \right. \\
& \quad \left. - 2^{-C} \int_0^C \int_0^{C-C_{s_{max}}} \dots \int_0^{C-C_{s_{max}}-\dots-C_3} C_2^3 \dots C_{s_{max}}^3 \text{poly}(C, C_2, \dots, C_{s_{max}}) dC_2 \dots dC_{s_{max}} \right. \\
& \quad \left. - \int_0^C \int_0^{C-C_{s_{max}}} \dots \int_0^{C-C_{s_{max}}-\dots-C_3} C_2^3 2^{-C_2} \dots \times C_{s_{max}}^3 2^{-C_{s_{max}}} (-F(0)) dC_2 \dots dC_{s_{max}} \right) \\
& = \text{Const.} \left((-F(0))^{s_{max}} - 2^{-C} \text{poly}(C) - (-F(0)) 2^{-C} \text{poly}(C) \right. \\
& \quad \left. - (-F(0))^2 \int_0^C \int_0^{C-C_{s_{max}}} \dots \int_0^{C-C_{s_{max}}-\dots-C_4} C_3^3 2^{-C_2} \dots \times C_{s_{max}}^3 2^{-C_{s_{max}}} dC_3 \dots dC_{s_{max}} \right) \\
& \quad \dots \\
& = \text{Const.} \left((-F(0))^{s_{max}} - \text{Const.} 2^{-C} \text{poly}(C) - (-F(0))^{s_{max}} \right) = \text{Const.} 2^{-C} \text{poly}(C).
\end{aligned}$$

Second inequality follows from independence of $[xx]_{e_1}, [xx]_{e_2}, \dots, [xx]_{e_{s_{max}}}$. Third inequality is due to the bound from proof 3.2.4.



Theorem 3.3.2. Taking $C = \log \alpha$, achievability result is established:

$$\mathbf{ARL2FA}(T_G) \geq \alpha; \mathbf{WADD}(T_G) \leq \frac{\log \alpha(1 + o(1))}{D(f_1||f_0)} \text{ as } \alpha \rightarrow \infty. \quad (3.15)$$

Therefore, proposed procedure T_G is asymptotically first-order instance optimal.

Proof. The first part of the theorem follows directly from theorem3.3.1. Optimality follows from comparing the converse result of corollary2.1.1 and the achievability result. \square

Notice that the procedure monitoring s_{max} edges is capable of detecting $s \leq s_{max}$ increased edges optimally. The price of monitoring more edges than necessary (when $s < s_{max}$) emerges at the higher-order terms of $\mathbf{ARL2FA}$ and the computational complexity of monitoring a larger amount of edges.

3.4 Representative Procedure for Edge Increased/Component Decreased

In section3.2, we proposed an optimal procedure based on monitoring every edge in $\mathcal{E}^{C;\text{forest}}$ in order to cover all possible locations for which the added edge should occur. In this section, we take advantage of the opportunity provided by correlated networks and devise a monitor efficient scheme requiring samples only from judiciously chosen nodes in the forest at the cost of some statistical performance.



3.4.1 Proposed Algorithm

$$T_{R-G} = \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}^R} \max_{\theta_{rs} \geq |\theta_{threshold}|} \sum_{i=k}^n \log(1 + \tanh(\theta_{rs}) x_r^{(i)} x_s^{(i)}) \geq C \right\}. \quad (3.16)$$

As in previous sections, we choose $\theta_{threshold} = \frac{1}{C}$. The only difference between this procedure and equation 3.3 lies in the edge set being monitored \mathcal{E}^R (call it the representative set), so the inner maximization problem is exactly the same as that of section 3.2.

The set \mathcal{E}^R is chosen such that for every two components in the forest, there exists one edge in \mathcal{E}^R connecting the two. Due to the construction, no matter where the increased edge appears, say that the increased edge (u, t) connects component one and two, there exists edge, say $(l, m) \in \mathcal{E}^R$, also connecting component one and two. Although in general $(l, m) \neq (u, t)$, node l will always be connected to m via $\text{path}(l, u) \cup \text{path}(u, t) \cup \text{path}(t, m)$. The ultimate effect is that although the precise location where the edge increased is not monitored, we will always measure two alternative nodes where there is a less intense change in correlation before and after the change due to the correlation propagation property.

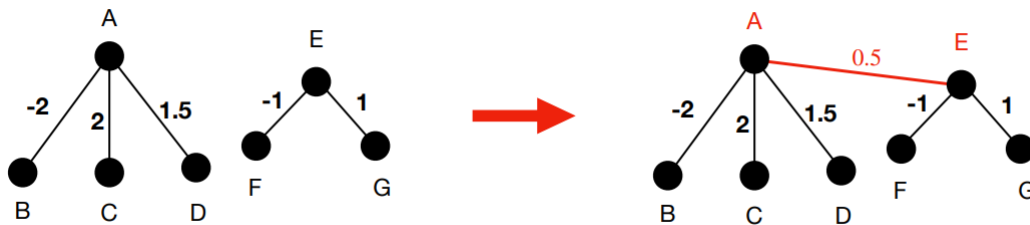


Figure 3.1: Illustration for one edge increase in Ising models on a forest.

Figures 3.1 and 3.2 better explains the intuition with visual aid. In figure 3.1, the

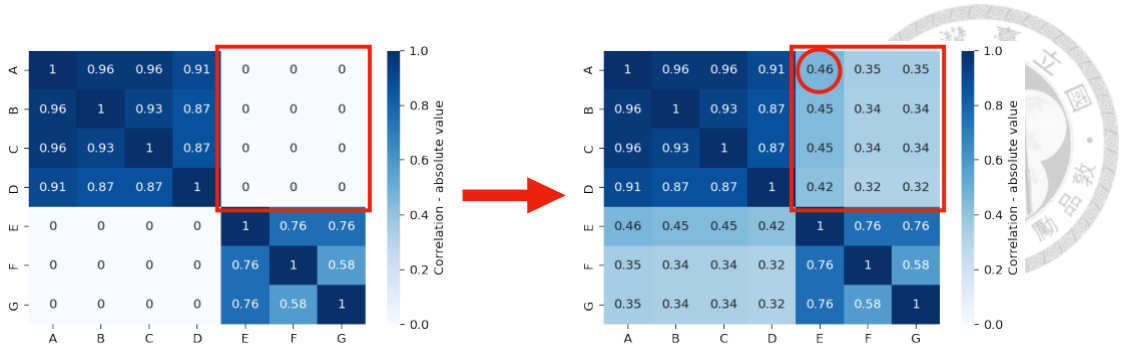


Figure 3.2: Illustration for absolute-valued correlation matrix.

distribution before the change is illustrated by the forest graph at the left. The numerical values are the weights θ for the zero mean-field Ising model on the forest. After the change, an additional edge appears, connecting nodes A and E with weight 0.5. Figure 3.2 shows the absolute-valued correlation matrix of the distribution before and after the change. The difference between the two is highlighted by the red rectangle. To detect the change, it suffices to monitor one among the twelve locations. However, monitoring any location other than the true change location (highlighted by the red circle) results in a larger delay due to the decay in correlation.

3.4.2 Performance for Proposed Algorithm

Denote using (u, t) the edge increased after the change and θ_{ut}^* the corresponding oracle edge weight. Due to the construction of \mathcal{E}^R , there exists some $(l, m) \in \mathcal{E}^R$ such that l belongs to the same component as u and m belongs to the same component as t .

Theorem 3.4.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$\mathbf{WADD}(T_{R-G}) \leq \frac{C(1 + o(1))}{1 - H\left(\frac{1 + \tanh(\theta_{ut}^*) \left(\prod_{\mu \in \text{path}(l, u)} \tanh(\theta_{\mu})\right) \left(\prod_{\mu \in \text{path}(m, t)} \tanh(\theta_{\mu})\right)}{2}\right)} \text{ as } C \rightarrow \infty. \quad (3.17)$$



$$\mathbf{ARL2FA}(T_{R-G}) \geq (1 + o(1))2^C. \quad (3.18)$$

Proof. To facilitate the proofs, define $\tau_{R-G} = \inf \{n \geq 1 : \max_{(r,s) \in \mathcal{E}^R} \max_{\theta_{rs} \geq |\theta_{\text{threshold}}|} \sum_{i=k}^n \log(1 + \tanh(\theta_{rs}) x_r^{(i)} x_s^{(i)}) \geq C\}$ and SPRT using the "marginal-correct" post-change parameter $\tau_{lm} = \inf \{n \geq 1 : \sum_{i=1}^n \log(1 + \tanh(\hat{\theta}_{lm}) x_l^{(i)} x_m^{(i)}) \geq C\}$. That is, $\tanh(\hat{\theta}_{lm}) = \tanh(\theta_{ut}^*) \left(\prod_{\mu \in \text{path}(l,u)} \tanh(\theta_\mu) \right) \left(\prod_{\mu \in \text{path}(m,t)} \tanh(\theta_\mu) \right)$.

Proof for WADD:

By Lorden's theorem, $\mathbf{WADD}(T_{R-G}) \leq \mathbb{E}_0[\tau_{R-G}]$. Furthermore,

$$\mathbb{E}_0[\tau_{R-G}] \leq \mathbb{E}_0[\tau_{lm}] \quad (3.19)$$

$$= \frac{C(1 + o(1))}{1 - H\left(\frac{1 + \tanh(\theta_{ut}^*) \left(\prod_{\mu \in \text{path}(l,u)} \tanh(\theta_\mu) \right) \left(\prod_{\mu \in \text{path}(m,t)} \tanh(\theta_\mu) \right)}{2}\right)} \text{ as } C \rightarrow \infty. \quad (3.20)$$

First inequality follows from the fact that representative generalized SPRT τ_{R-G} stops earlier than SPRT τ_{lm} . Final inequality derives from Wald's identity.

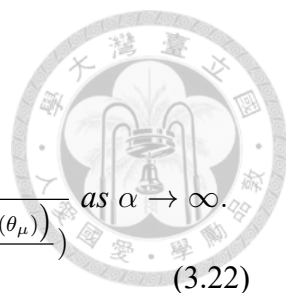
Proof for ARL2FA:

Use generalized CuSum lower bound to ARL2FA, theorem 3.2.1, and the fact that

$\mathbb{E}_\infty[T_{R-G}] \geq \mathbb{E}_\infty[T_G]$ concludes the proof. \square

Theorem 3.4.2. Taking $C = \log \alpha$, achievability results is established:

$$\mathbf{ARL2FA}(T_{R-G}) \geq \alpha; \quad (3.21)$$



$$\mathbf{WADD}(T_{R-G}) \leq \frac{\log \alpha(1 + o(1))}{1 - H\left(\frac{1 + \tanh(\theta_{ut}^*) \left(\prod_{\mu \in \text{path}(l,u)} \tanh(\theta_\mu)\right) \left(\prod_{\mu \in \text{path}(m,t)} \tanh(\theta_\mu)\right)}{2}\right)} \text{ as } \alpha \rightarrow \infty. \quad (3.22)$$

Therefore, proposed procedure T_{R-G} is not asymptotically first-order instance optimal.

Proof. The first part of the theorem follows directly from theorem3.4.1. Sub-Optimality follows from comparing the converse result of corollary2.1.1 and the achievability result. \square

Theorem3.4.2 shows that the procedure, as expected, loses some statistical performance:

$$D(f_1||f_0) = 1 - H\left(\frac{1 + \tanh(\theta_{ut}^*)}{2}\right) \quad (3.23)$$

$$\geq 1 - H\left(\frac{1 + \tanh(\theta_{ut}^*) \left(\prod_{\mu \in \text{path}(l,u)} \tanh(\theta_\mu)\right) \left(\prod_{\mu \in \text{path}(m,t)} \tanh(\theta_\mu)\right)}{2}\right). \quad (3.24)$$

3.4.3 Choosing the Representatives

In subsection3.4.1, we provided an algorithm monitoring pairs of nodes belonging to a representative set \mathcal{E}^R . We further required suitable conditions on the set so that every possible location for which the edge increased can be covered. In subsection3.4.2, we showed the trade-off between **WADD** and **ARL2FA**. Specifically in theorem3.4.2, we showed that the performance of the procedure not only depends on the oracle location (u, t) and parameter θ_{ut}^* but also on how we choose the representative set. A rational question to ask: How to choose a good representative set? Equivalently, the question is the same as asking for any two components \mathcal{T}_1 and \mathcal{T}_2 in the forest, how to select the pair

(l, m) with $l \in \mathcal{T}_1, m \in \mathcal{T}_2$?

According to theorem 3.4.2 and the fact that (u, t) is not known, a good representative set is the solution to the maximin problem:

$$\arg \max_{i \in \mathcal{T}_1, j \in \mathcal{T}_2} \min_{u \in \mathcal{T}_1, t \in \mathcal{T}_2} 1 - H\left(\frac{1 + \tanh(\theta_{ut}) \left(\prod_{\mu \in \text{path}(i, u)} \tanh(\theta_\mu)\right) \left(\prod_{\mu \in \text{path}(j, t)} \tanh(\theta_\mu)\right)}{2}\right). \quad (3.25)$$

$$\iff \arg \max_{i \in \mathcal{T}_1, j \in \mathcal{T}_2} \min_{u \in \mathcal{T}_1, t \in \mathcal{T}_2} \left| \tanh(\theta_{ut}) \left(\prod_{\mu \in \text{path}(i, u)} \tanh(\theta_\mu)\right) \left(\prod_{\mu \in \text{path}(j, t)} \tanh(\theta_\mu)\right) \right|. \quad (3.26)$$

$$\iff \arg \max_{i \in \mathcal{T}_1} \min_{u \in \mathcal{T}_1} \left| \prod_{\mu \in \text{path}(i, u)} \tanh(\theta_\mu) \right|; \arg \max_{j \in \mathcal{T}_2} \min_{t \in \mathcal{T}_2} \left| \prod_{\mu \in \text{path}(j, t)} \tanh(\theta_\mu) \right|. \quad (3.27)$$

The representative set selection problem reduces to a node selection problem for each component $\arg \max_{i \in \mathcal{T}_1} \min_{u \in \mathcal{T}_1} \prod_{\mu \in \text{path}(i, u)} \tanh(\theta_\mu)$. We provide a naive algorithm to solve the node selection problem.

1. $\forall i \in \mathcal{T}_1$, recursively solve $\min_{u \in \mathcal{T}_1} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)|$.

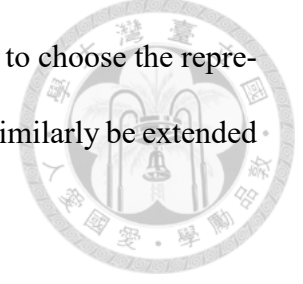
$$\begin{aligned} \min_{u \in \mathcal{T}_1} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)| &= \min_{u \in \text{leaves}(\mathcal{T}_1)} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)| \\ &= \min_{a \in N(i)} \left\{ \tanh(\theta_{ia}) \min_{u \in \text{leaves}(\mathcal{T}_{ai})} \prod_{\mu \in \text{path}(a, u)} |\tanh(\theta_\mu)| \right\} \end{aligned}$$

where $N(i)$ denotes the set of neighboring nodes of i , \mathcal{T}_{ai} is sub-tree of \mathcal{T}_1 constructed from breaking edge θ_{ai} and retaining the component with node a .

2. Take max over $i \in \mathcal{T}_1$.

When the pre-change weights are not known in advance, the algorithm is run with constant weight.

We finally remark that although we only showed explicitly how to choose the representative set under the problem setting of section 3.2, the method can similarly be extended to that of section 3.3.



3.5 Alternatives to Generalized Likelihoods for Shifts in Bernoulli

In section 3.2, we show that generalized CuSum finds the location for edge increase via brute force search and detects the change for unknown weight parameter via the Bernoulli generalized likelihood statistic. Both operations are computationally demanding, thereby motivating the need of procedures computationally more efficient. In the previous section, we replaced the brute force search of edge increase location by samples from only a few judiciously chosen nodes. In this section, we will give computationally simpler procedures replacing the Bernoulli generalized likelihood statistic.

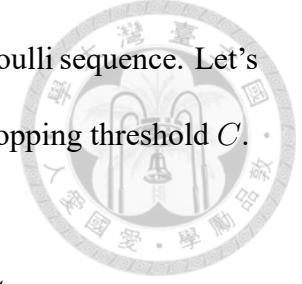
Let's first settle the Bernoulli quickest change detection problem setting.

1. Before the change, data sequence $\{x^{(1)}, x^{(2)}, \dots, x^{(\nu-1)}\}$ sampled i.i.d. from distribution $\mathbf{Ber}(\frac{1}{2})$.
2. After the unknown change point ν , data sequence $x^{(\nu)}, x^{(\nu+1)}, \dots$ sampled i.i.d from distribution $\mathbf{Ber}(\mu)$ with $\mu \neq \frac{1}{2}$.

3.5.1 Review of Bernoulli Generalized Likelihood Statistic

In subsection 3.1.2, by solving the inner maximization problem, we have derived the Bernoulli generalized likelihood statistic. Neglecting the presence of $\theta_{threshold}$, the statistic

is $n(1 - H(\frac{\text{count}_+}{n}))$, where count_+ denotes the number of 1 in the Bernoulli sequence. Let's get some intuition for the statistic by checking when it crosses the stopping threshold C .



$$n(1 - H(\frac{\text{count}_+}{n})) \geq C \Leftrightarrow H(\frac{\text{count}_+}{n}) \leq 1 - \frac{C}{n}. \quad (3.28)$$

Since $\frac{\text{count}_+}{n}$ is the ratio of ones (among the Bernoulli sequence) over the total number of samples, the statistic crosses C when the Bernoulli sequence observed is sufficiently unbalanced. Furthermore, the measure of balancedness shrinks with n . Without the shrinking for measure of balancedness or equivalently if the measure of balancedness remains fixed, there always exist some $\mu \neq \frac{1}{2}$ such that if data are sampled from $\mathbf{Ber}(\mu)$ the decision-maker wouldn't be able to detect and declare a change. Though shrinking the measure of balancedness is necessary, it should not be too severe; otherwise, under $\mathbf{Ber}(\frac{1}{2})$, the statistic would cross C too easily. It turns out the generalized likelihood ratio strikes a good balance.

Although generalized likelihood ratio has asymptotic first-order optimal performance, when using the statistic for quickest change detection procedures, the runtime complexity is high $\Theta(n^2)$.

$$T_G = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} (n - k + 1) (1 - H(\frac{\text{count}_+(k, n)}{n - k + 1})) \geq C \right\}. \quad (3.29)$$

3.5.2 Bernoulli Mixture Likelihood Statistic

Previously in definition 2.1.6, we have mentioned that the mixture approach is common in dealing with composite post-change distributions. Following the approach under the Bernoulli change detection problem and using uniform measure, we arrive at a mixture

likelihood ratio statistic highly related to the generalized likelihood ratio statistic.

$$\begin{aligned} \log \int 2^n \mu^{\text{count}_+} (1 - \mu)^{\text{count}_-} dG(\mu) &= n + \int_0^1 \mu^{\text{count}_+} (1 - \mu)^{\text{count}_-} d\mu \quad (3.30) \\ &= n + \log \frac{\text{count}_+! \text{count}_-!}{n!(n+1)} = n - \log\left(\binom{n}{\text{count}_+}\right) - \log(n+1). \quad (3.31) \end{aligned}$$

The integration is essentially the beta function. Note that the mixture statistic $n - \log\left(\binom{n}{\text{count}_+}\right) - \log(n+1)$ is asymptotically the same as the generalized statistic $n(1 - H(\frac{\text{count}_+}{n}))$ due to $\lim_{n \rightarrow \infty} \frac{\binom{n}{\text{count}_+}}{2^{nH(\frac{\text{count}_+}{n})}} = 1$.

Altogether we end up with a procedure with again high runtime complexity $\Theta(n^2)$.

$$T_M = \left\{ n \geq 1 : \max_{1 \leq k \leq n} ((n - k + 1) - \log\left(\binom{n - k + 1}{\text{count}_+(k, n)}\right) - \log(n - k + 2)) \geq C \right\}. \quad (3.32)$$

3.5.3 Consecutive Procedure

In the previous subsections, detection methods are based on variants of the likelihood ratio. In this subsection, we propose a procedure with lower runtime complexity explained by the simple intuition: When data samples are from $\text{Ber}(\frac{1}{2})$, it would be extremely rare to consecutively sample the same value. As a remark, we actually came up with the procedure by following ideas from Lorden[13]. Since their framework is fairly general and too complicated compared to what we have finally obtained, we will only explain our procedure.

Our procedure, denote using T_{con} , follows the rules:

1. First set a threshold C and initialize a counter $L_0 = 0$.

2. Take in Bernoulli samples. If the current sample differs from the previous sample, reset the counter $L_n = 0$; else, add one to the counter $L_n = L_{n-1} + 1$.
3. Once $L_n \geq \log C$, stop and declare a change.



The procedure essentially stops when $\log C$ consecutive sample take the same value. Note that the procedure runs in $\Theta(n)$, a significant improvement in runtime complexity. We quantify the performance of the procedure and show that the statistical performance is, however, greatly deteriorated.

Theorem 3.5.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$WADD(T_{con}) = \mathbb{E}_0[T_{con}] \leq \min\left\{\frac{C^{-\log \mu} - 1}{1 - \mu}, \frac{C^{-\log(1-\mu)}}{\mu}\right\}. \quad (3.33)$$

$$ARL2FA(T_{con}) \geq C. \quad (3.34)$$

Proof. Proof for WADD:

The first equality follows since T_{con} performs exactly the same regardless of the change point time ν and past data $\{x_i\}_{i=1, \dots, \nu-1}$. The inequality follows the idea that the expected stopping time is less than or equal to both the expected number of tosses until $\log A$ consecutive 1 tosses and the expected number of tosses until $\log A$ consecutive 0 tosses. Define E_n to be the number of tosses until n consecutive 1 tosses, with 1/0 tosses happening with probability $\mu/1 - \mu$ respectively and $\{x_i\}_i$ be the sequence of

tosses. Then,

$$\mathbb{E}[E_n] = (1 - \mu)\mathbb{E}[E_n|x_1 = 0] + \mu(1 - \mu)\mathbb{E}[E_n|x_1 = 1, x_2 = 0] + \dots \quad (3.35)$$

$$+ \mu^{n-1}(1 - \mu)\mathbb{E}[E_n|x_1, \dots, x_{n-1} = 1, x_n = 0] + \mu^n\mathbb{E}[E_n|x_1, \dots, x_n = 1] \quad (3.36)$$

$$= (1 - \mu)(\mathbb{E}[E_n] + 1) + \mu(1 - \mu)(\mathbb{E}[E_n] + 2) + \dots + \mu^{n-1}(1 - \mu)(\mathbb{E}[E_n] + n) + \mu^n n \quad (3.37)$$

$$\Leftrightarrow \mathbb{E}[E_n] = \frac{\frac{1}{\mu^n} - 1}{1 - \mu}. \quad (3.38)$$

Plug in $n = \log C$ and follow similar analysis for consecutive 0 tosses concludes the proof for WADD.

Proof for ARL2FA:

Proof follows similar idea from the analysis of WADD proof. Similarly define $E_{n,1}, E_{n,0}$ to be respectively the number of tosses until n consecutive 1, 0 tosses. Under $\mathbf{Ber}(\frac{1}{2})$,

$$\mathbb{E}[T_{con}] = \frac{1}{2}\mathbb{E}[E_{n,1}] + \frac{1}{2}\mathbb{E}[E_{n,0}] = 2 \times \frac{1}{2} \times \frac{(\frac{1}{2})^{-n} - 1}{1 - \frac{1}{2}} = 2(2^n - 1). \quad (3.39)$$

Plug in $n = \log C$. □

Theorem 3.5.2. Taking $C = \alpha$, achievability result is established:

$$\mathbf{ARL2FA}(T_G) \geq \alpha; \quad \mathbf{WADD}(T_G) \leq \min\left\{\frac{\alpha^{-\log \mu} - 1}{1 - \mu}, \frac{\alpha^{-\log(1-\mu)}}{\mu}\right\}. \quad (3.40)$$

Therefore, consecutive procedure is statistically worse than generalized CuSum.

Proof. The first part of the theorem follows directly from theorem3.5.1. Sub-optimality is evident by comparing with the result from theorem3.2.2. □

3.5.4 Fixed Window Procedure



It's not surprising that the consecutive procedure from the previous subsection performs poorly as we already know from the generalized likelihood ratio statistic, the measure of balancedness has a critical role to play in the Bernoulli change detection problem. The consecutive procedure measures balancedness simply too crudely. In this subsection, we modify generalized CuSum to a window-fixed procedure and analyze its performance.

Recall that the generalized CuSum procedure requires runtime complexity $\Theta(n^2)$. As compared to the simple CuSum procedure, an additional multiple of n is required due to the presence of $\max_{1 \leq k \leq n}$ and the fact that generalized ratio statistic no longer possess the recursive structure. The operation $\max_{1 \leq k \leq n}$ can be viewed as scanning over all window sizes, i.e., sequences monitored at time n have window sizes of $1, 2, \dots, n - 1$. To reduce the runtime complexity, we replace $\max_{1 \leq k \leq n}$ with $k = n - w$ where w is a pre-chosen window size.

$$T_w = \inf \left\{ n \geq 1 : w(1 - H(\frac{\text{count}_+(n - w + 1, n)}{w})) \geq C \right\}. \quad (3.41)$$

Clearly, the performance of the fixed window procedure depends on the window size. We will parameterize w by δ and choose $w = \frac{C}{D(\delta || \frac{1}{2})}$. Here $D(\delta || \frac{1}{2})$ is the short-hand notation for the KL-divergence between Bernoulli δ and Bernoulli $\frac{1}{2}$.

Theorem 3.5.3. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

When window size is chosen with $|\delta - \frac{1}{2}| \geq |\mu - \frac{1}{2}|$,

$$\mathbf{WADD}(T_w) \leq \frac{C}{D(\delta||\frac{1}{2})} \sqrt{\frac{2C}{D(\delta||\frac{1}{2})}} \exp\left\{\frac{D(\delta||\mu)}{D(\delta||\frac{1}{2})}C\right\} \quad (3.42)$$



Otherwise,

$$\mathbf{WADD}(T_w) \leq \frac{C}{D(\delta||\frac{1}{2})} \sqrt{\frac{2C}{D(\delta||\frac{1}{2})}}. \quad (3.43)$$

For any window size,

$$\mathbf{ARL2FA}(T_w) \geq 2^C - \frac{C}{D(\delta||\frac{1}{2})}. \quad (3.44)$$

Proof. Proof for WADD:

Define the event $E_n = \{w(1 - H(\frac{\text{count}_+(n, n+w)}{w})) < C\}$.

$$\mathbb{E}_0[T_w] = \sum_{t=0}^{\infty} P_0(T_w > t) = \sum_{t=0}^w P_0(T_w > t) + \sum_{t=w}^{2w} P_0(T_w > t) + \sum_{t=2w}^{3w} P_0(T_w > t) + \dots \quad (3.45)$$

$$\leq \sum_{t=0}^w 1 + \sum_{t=w}^{2w} P_0(E_w) + \sum_{t=2w}^{3w} P_0(E_w \cap E_{2w}) + \dots \quad (3.46)$$

If $|\delta - \frac{1}{2}| \geq |\mu - \frac{1}{2}|$, then by anti-concentration for binomial distribution:

$$P_0(E_w) \geq \frac{1}{\sqrt{2w}} \exp\{-wD(\delta||\mu)\}. \quad (3.47)$$

Thus,

$$\mathbb{E}_0[T_w] \leq w(1 + (1 - \frac{1}{\sqrt{2w}} \exp\{-wD(\delta||\mu)\}) + (1 - \frac{1}{\sqrt{2w}} \exp\{-wD(\delta||\mu)\})^2 + \dots) \quad (3.48)$$

$$= w\sqrt{2w} \exp\{wD(\delta||\mu)\} = \frac{C}{D(\delta||\frac{1}{2})} \sqrt{\frac{2C}{D(\delta||\frac{1}{2})}} \exp\left\{\frac{D(\delta||\mu)}{D(\delta||\frac{1}{2})}C\right\} \quad (3.49)$$

Otherwise if $|\delta - \frac{1}{2}| \leq |\mu - \frac{1}{2}|$, anti-concentration for binomial distribution:

$$P_0(E_w) \geq \frac{1}{\sqrt{2w}} \exp\{-wD(\delta||\delta)\} = \frac{1}{\sqrt{2w}} \quad (3.50)$$



Thus,

$$\mathbb{E}_0[N] \leq w\sqrt{2w} = \frac{C}{D(\delta||\frac{1}{2})} \sqrt{\frac{2C}{D(\delta||\frac{1}{2})}} \quad (3.51)$$

Proof for ARL2FA:

Denote using T_G the generalized CuSum procedure.

$$\mathbb{E}_\infty[T_w] = \mathbb{E}_\infty[T_w | T_w \geq w] \geq \mathbb{E}_\infty[T_G | T_G \geq w] \geq P_\infty(T \geq w) \mathbb{E}_\infty[T_G | T_G \geq w] \quad (3.52)$$

$$\geq 2^C - P_\infty(T_G < w) \mathbb{E}_\infty[T_G | T_G < w] \geq 2^C - 1 \times w. \quad (3.53)$$

□

Theorem 3.5.4. Taking $C = \log \alpha$, achievability result is established:

$$\mathbf{ARL2FA}(T_w) \geq \alpha. \quad (3.54)$$

When window size is chosen with $|\delta - \frac{1}{2}| \geq |\mu - \frac{1}{2}|$,

$$\mathbf{WADD}(T_w) \leq \frac{\log \alpha}{D(\delta||\frac{1}{2})} \sqrt{\frac{2 \log \alpha}{D(\delta||\frac{1}{2})}} \exp\left\{\frac{D(\delta||\mu)}{D(\delta||\frac{1}{2})} \log \alpha\right\}. \quad (3.55)$$

Otherwise,

$$\mathbf{WADD}(T_w) \leq \frac{\log \alpha}{D(\delta||\frac{1}{2})} \sqrt{\frac{2 \log \alpha}{D(\delta||\frac{1}{2})}}. \quad (3.56)$$

Therefore, the fixed window procedure performs better than the consecutive procedure but is of course worse than generalized CuSum.



Proof. The first part of the theorem follows directly from theorem 3.5.3. \square

3.6 One Edge Increase in General Ising Models and Implications

Previously in section 3.2, we showed that the one edge increased structural changing Ising model on a forest problem is closely related to the correlation Bernoulli mean shift problem. In this section, we will show similar conclusion for the one edge increased structural changing general Ising model problem.

3.6.1 Problem Setting

Under the quickest change detection problem setting, we further assume that:

1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from distribution f_0 belonging to general Ising models (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{Z} \exp\{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t\}$
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution f_1 differing from pre-change distribution by one edge **more**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{Z} \exp\{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}^+} \theta_{ut} x_u x_t\}$).

Let us first assume that parameters before the change δ_s, θ_{ut} are known.

3.6.2 Extended Algorithm from Subsection 3.2.2 and its Performance



When applying generalized CuSum to the general Ising model, the mere difference from that of subsection 3.2.2 is the likelihood ratio.

$$T_G = \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}^G} \max_{\theta_{rs} \geq |\theta_{t \text{ threshold}}|} \sum_{i=k}^n \log \left(\frac{A_{rs} + B_{rs}}{e^{\theta_{rs}} A_{rs} + e^{-\theta_{rs}} B_{rs}} e^{\theta_{rs} x_r^{(i)} x_s^{(i)}} \right) \geq C \right\}. \quad (3.57)$$

$$\text{where } A_{rs} = \sum_{\{\mathbf{x}: x_r x_s = +1\}} e^{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t} \text{ and } B_{rs} = \sum_{\{\mathbf{x}: x_r x_s = -1\}} e^{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t}.$$

We will show that after examining a particular edge (r, s) via the maximum operation, the generalized likelihood ratio of the general Ising model is exactly the same as that of generalized likelihood ratio for Bernoulli shifts. The generalized Bernoulli likelihood ratio for detecting shift from $\mathbf{Ber}(\theta)$ to $\mathbf{Ber}(\theta')$ in Bernoulli sequence:

$$\max_{\theta'} \left(\frac{\theta'}{\theta} \right)^{\text{count}_+} \left(\frac{1 - \theta'}{1 - \theta} \right)^{\text{count}_-}. \quad (3.58)$$

The notation count_+ , count_- is respectively the number of 1 and 0 in the Bernoulli sequence.

The generalized likelihood ratio for general Ising model targeted to edge (r, s) :

$$\max_{\theta_{rs}} \left(\frac{e^{\theta_{rs}} (A_{rs} + B_{rs})}{e^{\theta_{rs}} A_{rs} + e^{-\theta_{rs}} B_{rs}} \right)^{\text{count}_{+;rs}} \left(\frac{e^{-\theta_{rs}} (A_{rs} + B_{rs})}{e^{\theta_{rs}} A_{rs} + e^{-\theta_{rs}} B_{rs}} \right)^{\text{count}_{-;rs}}. \quad (3.59)$$

Comparing the generalized likelihood ratios and let $\theta = \frac{A_{rs}}{A_{rs} + B_{rs}}$ and $\theta' = \frac{e^{\theta_{rs}} A_{rs}}{e^{\theta_{rs}} A_{rs} + e^{-\theta_{rs}} B_{rs}}$, the latter problem is equivalent to the former problem. When the pre-change weights are known to the decision-maker, the quantities A_{rs} , B_{rs} can be computed and the problem es-

entially becomes a Bernoulli quickest change detection problem with known pre-change and unknown post-change distribution. Therefore, following similar analysis from subsection 3.2.3 and subsection 3.2.4 generalized CuSum can be shown to be asymptotically first-order optimal.

Theorem 3.6.1. Taking $C = \log \alpha$,

$$\mathbf{ARL2FA}(T_G) \geq \alpha; \mathbf{WADD}(T_G) \leq \frac{\log \alpha(1 + o(1))}{D(f_1 || f_0)} \text{ as } \alpha \rightarrow \infty. \quad (3.60)$$

Proof. Refer to the proof for theorem 3.2.2. □

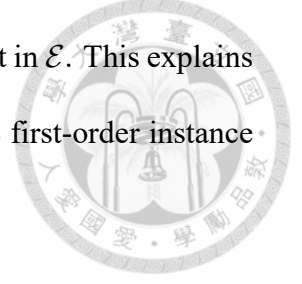
3.6.3 Implications

In the previous subsection, we argued that with the aid of pre-change parameters, detecting an increased edge in general Ising models is closely related to the correlation Bernoulli mean shift problem (from $\mathbf{Ber}(\frac{A_{rs}}{A_{rs}+B_{rs}})$ to $\mathbf{Ber}(\mu)$; $\mu \neq \frac{A_{rs}}{A_{rs}+B_{rs}}$), whereas in subsection 3.1.2, we have mentioned that the same problem in Ising forests is closely related to the correlation $\mathbf{Ber}(\frac{1}{2})$ mean shift problem. The goal of this subsection is to provide a connection between the two settings, explain why we were able to achieve first-order asymptotic instance optimality with only the structural information in section 3.2, and extend the instance optimality to slightly more general models.

Since Ising models on a forest is simply a sub-family of the general Ising model, we should be able to deduce the results of section 3.2 from results under the general model. In fact, the deduction is simple: $\mathbf{Ber}(\frac{A_{rs}}{A_{rs}+B_{rs}})$ reduces to $\mathbf{Ber}(\frac{1}{2})$ when $A_{rs} = B_{rs}$ for all $(r, s) \in \mathcal{E}^{C;\text{forest}}$. When the underlying distribution is confined to zero mean-field Ising

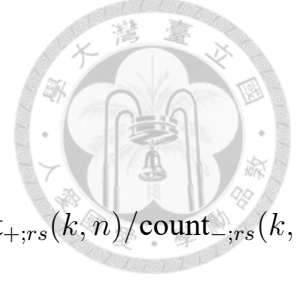
models on a forest, $A_{rs} = \sum_{\{\mathbf{x}: x_r x_s = +1\}} \prod_{(u,t) \in \mathcal{E}} (1 + \tanh(\theta_{ut} x_u x_t)) = B_{rs} = \sum_{\{\mathbf{x}: x_r x_s = -1\}} \prod_{(u,t) \in \mathcal{E}} (1 +$

$\tanh(\theta_{ut}x_u x_t)$ regardless of the pre-change weights, since (r, s) is not in \mathcal{E} . This explains why the pre-change weights are not necessary to achieve asymptotic first-order instance optimality.



Arguments from the previous paragraph naturally raises the question: Does there exist a class of distribution broader than Ising models on a forest such that $\mathbf{Ber}(\frac{A_{rs}}{A_{rs}+B_{rs}})$ also reduces to $\mathbf{Ber}(\frac{1}{2})$? If so, asymptotic first-order optimality for the one edge increase quickest change detection problem on such models can be achieved using the same procedure from subsection 3.2.2; structural information is sufficient again! Here we give an example. A broader class of Ising models equipped with the aforementioned property is Ising models with zero mean-field vector $\delta_s = 0 \forall s \in \mathcal{V}$ and multiple disconnected components. Furthermore, only bridges are allowed appear after the change. The setting intuitively reduces to the correlation $\mathbf{Ber}(\frac{1}{2})$ mean shift problem since before the change each node has equal marginal probability (due to zero mean-field vector) and disconnectedness renders independence for x_r and x_s . After the change, the non-zero weight θ_{rs} breaks the independence and brings a bias to $x_r x_s$. The reduction is mathematically proven by $A_{rs} = \sum_{\{\mathbf{x}: x_r x_s = +1\}} e^{\sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t} = B_{rs} = \sum_{\{\mathbf{x}: x_r x_s = -1\}} e^{\sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t}$. This class of model is more general than forests in the sense that disjointed components are allowed to have loops.

3.7 Some Lemmas



Proof for Inner Maximization. For notational simplicity, denote $\text{count}_{+,rs}(k, n)/\text{count}_{-,rs}(k, n)$ using $\text{count}_+/\text{count}_-$.

$$f(\theta_{rs}) := \prod_{i=k}^n (1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}) = ((1 + \tanh \theta_{rs})(1 - \tanh \theta_{rs}))^{\text{count}_-} (1 + \tanh \theta_{rs})^{\text{count}_+ - \text{count}_-} \quad (3.61)$$

$$= (\text{sech } \theta_{rs})^{2\text{count}_-} (1 + \tanh \theta_{rs})^{\text{count}_+ - \text{count}_-}. \quad (3.62)$$

$$f'(\theta_{rs}) = (\text{sech } \theta_{rs})^{(2\text{count}_+)} (1 + \tanh \theta_{rs})^{(\text{count}_+ - \text{count}_- - 1)} \quad (3.63)$$

$$((\text{count}_+ - \text{count}_-)(\text{sech } \theta_{rs})^2 + (-2\text{count}_- \tanh \theta_{rs})(1 + \tanh \theta_{rs})). \quad (3.64)$$

$$f'(\theta_{rs}^*) = 0 \text{ yields } (\text{count}_+ - \text{count}_-)(\text{sech } \theta_{rs}^*)^2 - 2\text{count}_- \tanh \theta_{rs}^* (1 + \tanh \theta_{rs}^*) = 0.$$

Further calculations:

$$\frac{\text{count}_+ - \text{count}_-}{2\text{count}_-} = \frac{\tanh \theta_{rs}^* (1 + \tanh \theta_{rs}^*)}{(\text{sech } \theta_{rs}^*)^2} = \frac{e^{\theta_{rs}^*} (e^{\theta_{rs}^*} - e^{-\theta_{rs}^*})}{2} = \frac{e^{2\theta_{rs}^*} - 1}{2}.$$

Finally we have,

$$\theta_{rs}^* = \frac{1}{2} \ln\left(\frac{\text{count}_+ - \text{count}_-}{\text{count}_-} + 1\right) = \frac{1}{2} \ln\left(\frac{\text{count}_+}{\text{count}_-}\right). \quad (3.65)$$

□



Chapter 4 QCD for Edge Decreased in Ising Models on a Forest

The goal of this chapter is to study the complementary type of structural change in Ising models on a forest: edge disappearance. Depending on the decision-maker's knowledge of the pre/post-change distributions, the detection problem has a distinct degree of difficulty. The first and simplest scenario we study in this chapter is when the decision-maker has sufficient confidence and samples from the pre-change distribution. These samples can be used to derive estimates of the pre-change distributions, thereby easing future quickest change detection tasks. We characterize the fundamental limit in this setting by showing generalized CuSum achieves asymptotic first-order optimality. Another scenario where only the sign and support of the pre-change distribution is known to the decision-maker is next studied. We provide an algorithm that is no longer instance optimal but is optimal in the sense of [14]. Next, we study how to take advantage of the correlation propagation characteristic of the model and further provide a more efficient detection procedure at the cost of some statistical performance. Finally, we conclude with preliminary results for the detection of one edge decrease in general Ising models and discuss its implications.



4.1 One Edge Decreased/One Component Increased with Known Pre-change Distribution

4.1.1 Problem Setting

Under the quickest change detection problem setting, we further assume that:

1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from known distribution f_0 belonging to Ising model on forest (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j)$).
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution f_1 differing from pre-change distribution by one edge **less**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}^-} (1 + \tanh(\theta_{ij})x_i x_j)$).

In this setting, note that before the change, edge set \mathcal{E} and parameters $\theta_{ij} \forall \{i, j\} \in \mathcal{E}$ are completely specified. After the change, edge set $\mathcal{E}^- = \mathcal{E} \setminus \{r, s\}$ for some $\{r, s\} \in \mathcal{E}$; parameters $\theta_{ij} \in \mathcal{E} \setminus \{r, s\}$ remains untouched.

4.1.2 Proposed Algorithm

Since under this setting, the problem belongs to a composite post-change distribution quickest change detection problem, the generalized or mixture approach may work decently. Following the generalized approach, we end up with:

$$T_G = \inf\{n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}} \sum_{i=k}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} \geq C\}. \quad (4.1)$$

Intuitively, the procedure locates the location of the decreased edge by taking a maximum operation over all edges presented in the tree. The procedure has computational complexity equal to $\Theta(|\mathcal{E}|n)$ and can be written in a recursive form.



$$W_{rs}^{(1)} = 0; W_{rs}^{(n+1)} = (W_{rs}^{(n)} + \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(n+1)} x_s^{(n+1)}}, 0)^+ \quad \forall (r, s) \in \mathcal{E}. \quad (4.2)$$

$$T_{G,rs} = \inf\{n \geq 1 : W_{rs}^{(n)} \geq C\}; T_G = \min_{(r,s) \in \mathcal{E}} \{T_{G,rs}\}. \quad (4.3)$$

4.1.3 Performance of Algorithm

We evaluate the proposed algorithm under Lorden's criteria by studying the algorithm's worst average detection delay and average run length to false alarm.

Theorem 4.1.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$\mathbf{WADD}(T_G) \leq \frac{C(1 + o(1))}{D(f_1 || f_0)} \text{ as } C \rightarrow \infty. \quad (4.4)$$

$$\mathbf{ARL2FA}(T_G) \geq \frac{2^C}{|\mathcal{E}|}. \quad (4.5)$$

Proof. To facilitate the proofs, define $\tau_G = \inf\{n \geq 1 : \max_{(r,s) \in \mathcal{E}} \sum_{i=1}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} \geq C\}$ and $\tau_{rs} = \inf\{n \geq 1 : \sum_{i=1}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} \geq C\}$.

Proof for WADD:

By Lorden's theorem 2.1.3, $\mathbf{WADD}(T_G) \leq \mathbb{E}_0[\tau_G]$. Denote using (u, t) the real edge

deleted, then

$$\mathbb{E}_0[\tau_G] \leq \mathbb{E}_0[\tau_{ut}] = \frac{C(1 + o(1))}{D(f_1||f_0)} \text{ as } C \rightarrow \infty.$$

First inequality follows from the fact that τ_G stops earlier than τ_{ut} . Final inequality derives from Wald's identity.

Proof for ARL2FA:

By Lorden's theorem 2.1.3, to prove $\mathbf{ARL2FA}(T_G) \geq \frac{2^C}{|\mathcal{E}|}$, it suffice to show $P_\infty(\tau_G < \infty) \leq \frac{|\mathcal{E}|}{2^C}$.

$$\begin{aligned} P_\infty(\tau_G < \infty) &= \lim_{t \rightarrow \infty} P_\infty(\tau_G < t) = \lim_{t \rightarrow \infty} P_\infty(\cup_{(r,s) \in \mathcal{E}} \{\tau_{rs} < t\}) \\ &\leq \sum_{(r,s) \in \mathcal{E}} \lim_{t \rightarrow \infty} P_\infty(\tau_{rs} < t) \leq \frac{|\mathcal{E}|}{2^C} \end{aligned}$$

Final inequality follows from maximal inequality. □

Theorem 4.1.2. *Taking $C = \log \alpha$, achievability results is established:*

$$\mathbf{ARL2FA}(T_G) \geq \alpha; \mathbf{WADD}(T_G) \leq \frac{\log \alpha(1 + o(1))}{D(f_1||f_0)} \text{ as } \alpha \rightarrow \infty. \quad (4.6)$$

Therefore, T_G is, not surprising, asymptotic first-order optimal.

Proof. The first part of the theorem follows directly from theorem 4.1.1. Optimality follows from comparing the converse result of corollary 2.1.1 and the achievability result. □



4.2 One Edge Decreased/One Component Increased with only Sign of Pre-change Distribution



4.2.1 Problem Setting

In this section, the problem setting is almost the same as that of the previous section. The only difference is that here we assume that the pre-change distribution is not completely known to the decision-maker; only the sign, support, and interval for which the parameters reside is known. The assumption is reasonable since side information from the pre-change or regular state should be able to tell between the following situations: cooperative, opposing, or not related. Mathematically stated, we assume that:

1. Before the change, data sequence $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\nu-1)}\}$ sampled i.i.d. from distribution f_0 belonging to Ising model on forest (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}} (1 + \tanh(\theta_{ij})x_i x_j)$). **Parameters θ_{ij} are not known but sign of parameters $\text{sgn}(\theta_{ij})$, upper/lower bounds to parameters $\theta_-, \theta_+ > 0$ ($0 < \theta_- \leq \theta_{ij} \leq \theta_+ < \infty$ if $\theta_{ij} > 0$, $-\infty < -\theta_+ \leq \theta_{ij} \leq -\theta_- < 0$ otherwise) are known, and support \mathcal{E} is known.**
2. After the unknown change point ν , data sequence $\{\mathbf{x}^{(\nu)}, \mathbf{x}^{(\nu+1)}, \dots\}$ sampled i.i.d from distribution P_1 differing from pre-change distribution by one edge **less**. (i.e. pmf for samples $f(\mathbf{x}) = \frac{1}{2^d} \prod_{(i,j) \in \mathcal{E}^-} (1 + \tanh(\theta_{ij})x_i x_j)$).

4.2.2 Proposed Algorithm



The likelihood ratio $\frac{f_1(x)}{f_0(x)} = \frac{1}{1 + \tanh \theta_{ut} x_u x_t} = \frac{\frac{1}{2}}{\frac{1 + \tanh \theta_{ut} x_u x_t}{2}}$ hints that the problem is closely related to the pre-change unknown Bernoulli quickest change detection problem (i.e. shifting from $\mathbf{Ber}(\frac{1 + \tanh \theta_{ut}}{2})$ to $\mathbf{Ber}(\frac{1}{2})$). Therefore, we modify Mei's algorithm for one-parameter exponential families by adding a maximum operation to locate the deleted edge and derive the algorithm:

$$T_D = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}} \min_{\theta_{rs}} \left\{ \sum_{i=k}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} - \log(\cosh \theta_{rs}) C \right\} \geq 0 \right\}. \quad (4.7)$$

Minimization domain of θ_{rs} depends on its sign; if $\theta_{rs} < 0$, minimization is taken over $-\infty < -\theta_+ \leq \theta_{rs} \leq -\theta_- < 0$; otherwise $0 < \theta_- \leq \theta_{rs} \leq \theta_+ < \infty$.

The inner minimization over θ_{rs} has a simple structure, which can be used to further simplify the expression of T_D .

$$\sum_{i=1}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} \geq \log(\cosh \theta) C \text{ for all } \mathbf{sgn}(\theta) \theta_+ < \theta < \mathbf{sgn}(\theta) \theta_- \quad (4.8)$$

$$\iff \sum_{i=1}^n \log \frac{e^{\theta x_r^{(i)} x_s^{(i)}} + e^{-\theta x_r^{(i)} x_s^{(i)}}}{2e^{\theta x_r^{(i)} x_s^{(i)}}} \geq \log(\cosh \theta) C \text{ for all } \mathbf{sgn}(\theta) \theta_+ < \theta < \mathbf{sgn}(\theta) \theta_- \quad (4.9)$$

$$\iff \theta(-\text{count}_+ + \text{count}_-) + n \log(\cosh \theta) \geq C \log(\cosh \theta) \text{ for all } \mathbf{sgn} \theta_+ < \theta < \mathbf{sgn} \theta_- \quad (4.10)$$

If $\theta < 0$,

$$-\text{count}_+ + \text{count}_- \leq (C - n) \frac{\log(\cosh \theta)}{\theta} \text{ for all } -\theta_+ < \theta < -\theta_- \quad (4.11)$$



When $C > n$, minimum attains at $-\theta_+$; when $C \leq n$, minimum attains at $-\theta_-$.

If $\theta \geq 0$,

$$-\text{count}_+ + \text{count}_- \geq (C - n) \frac{\log(\cosh \theta)}{\theta} \text{ for all } \theta_- < \theta < \theta_+ \quad (4.12)$$

When $C > n$, maximum attains at θ_+ ; when $C \leq n$, maximum attains at θ_- .

Therefore, the equivalent expression as suggested by Mei:

$$T_D = \inf\{n \geq 1 : \max_{n-C+1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}} \sum_{i=k}^n \log\left(\frac{1}{1 + \tanh(\mathbf{sgn}(\theta_{rs})\theta_-)x_r^{(i)}x_s^{(i)}}\right)\} \quad (4.13)$$

$$\geq \log(\cosh(\mathbf{sgn}(\theta_{rs})\theta_-))C \text{ or} \quad (4.14)$$

$$\max_{(r,s) \in \mathcal{E}} W_{n-C;(r,s)} + \sum_{i=n-C+1}^n \log\left(\frac{1}{1 + \tanh(\mathbf{sgn}(\theta_{rs})\theta_+)x_r^{(i)}x_s^{(i)}}\right) \quad (4.15)$$

$$\geq \log(\cosh(\mathbf{sgn}(\theta_{rs})\theta_+))C\}. \quad (4.16)$$

where $W_{k;(r,s)} = (W_{k-1;(r,s)} + \log(\frac{1}{1 + \tanh(\mathbf{sgn}(\theta_{rs})\theta_+)x_r^{(i)}x_s^{(i)}}))^{+}$.

4.2.3 Performance of Algorithm

The proposed algorithm is evaluated under the reverse of Lorden's criteria by studying the algorithm's worst average detection delay and average run length to false alarm.

Theorem 4.2.1. *Upper bound to worst average detection delay and lower bound to aver-*

age run length to false alarm:

$$\mathbf{WADD}(T_D) \leq C(1 + o(1)) \text{ as } C \rightarrow \infty.$$



$$\mathbf{ARL2FA}(T_D) \geq \frac{2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))C}}{|\mathcal{E}|} = \frac{(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))^C}{|\mathcal{E}|}. \quad (4.18)$$

Proof for WADD. To facilitate the proofs, define $\tau_D = \inf \{n \geq 1 : \max_{(r,s) \in \mathcal{E}} \min_{\theta_{rs}} \{\sum_{i=k}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} - \log(\cosh \theta_{rs})C\} \geq 0\}$ and $\tau_{rs} = \inf \{n \geq 1 : \min_{\theta_{rs}} \{\sum_{i=k}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} - \log(\cosh \theta_{rs})C\} \geq 0\}$.

Proof for WADD:

By Lorden's theorem 2.1.3, $\mathbf{WADD}(T_D) \leq \mathbb{E}_0[\tau_D]$. Denote using (u, t) the real edge deleted, then

$$\mathbb{E}_0[\tau_D] \leq \mathbb{E}_0[\tau_{ut}] = C(1 + o(1)) \text{ as } C \rightarrow \infty.$$

Equality derives from lemma 4.5.1 and with exponential family restricted to pmf $f(x) = \exp\{\gamma x - b(\gamma)\}$, $x \in \{-1, +1\}$, and $b(\gamma) = \log(\cosh \gamma)$. In our problem, we would like to detect a shift from parameter γ to 0. Therefore, $b''(\gamma = 0) = \text{sech}^2(\gamma)|_{\gamma=0} = 1$ and plugging in the kl-divergences, we get

$$\mathbb{E}_0[\tau_{ut}] = C + \frac{1}{\sqrt{2\pi}} \left(\frac{\theta_-}{\log(\cosh \theta_-)} - \frac{\theta_+}{\log(\cosh \theta_+)} + o(1) \right) \sqrt{C} \text{ as } C \rightarrow \infty. \quad (4.19)$$

Proof for ARL2FA:

By Lorden's theorem 2.1.3, to prove $\mathbf{ARL2FA}(T_D) \geq \frac{2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))C}}{|\mathcal{E}|}$, it suffice to

show $P_\infty(\tau_D < \infty) \leq \frac{|\mathcal{E}|}{2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))C}}.$

$$P_\infty(\tau_D < \infty) = \lim_{t \rightarrow \infty} P_\infty(\cup_{(r,s) \in \mathcal{E}} \{\tau_{rs} < t\}) \quad (4.20)$$

$$\leq \sum_{(r,s) \in \mathcal{E}} \lim_{t \rightarrow \infty} P_\infty(\tau_{rs} < t) \leq \sum_{(r,s) \in \mathcal{E}} \frac{1}{2^{\log(\cosh \theta_{rs})C}} \leq \frac{|\mathcal{E}|}{2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))C}}. \quad (4.21)$$

Second inequality follows from the fact that τ_{rs} stops later than sequential probability ratio test with true pre-change parameter. \square

Theorem 4.2.2. *With threshold $C = \alpha$, achievability result is established:*

$$\mathbf{WADD}(T_D) \leq \alpha; \mathbf{ARL2FA}(T_D) \geq 2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))\alpha} = (\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))^\alpha \text{ as } \alpha \rightarrow \infty. \quad (4.22)$$

Therefore, T_D is not asymptotically first-order optimal under reverse Lorden's criteria.

Proof. The first part of the theorem follows directly from theorem4.2.1. Sub-Optimality follows from comparing the converse result of corollary2.1.1 and the achievability result. \square

Since the performance of the procedure is dominated by the pre-change edge with smallest absolute value and that the true kl-divergence $D(f_1||f_0) \geq \log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|))$, the procedure is not instance optimal. However, it's optimal in the sense of definition2.1.11; for every pre-change weight, T_D is asymptotic efficient (see definition2.1.10) when the decreased edge corresponds to the one with the smallest absolute value; for every decreased edge, T_D is asymptotic efficient when the pre-change weights satisfy the condition that the deleted edge has the smallest absolute value.

4.3 Representative Procedure for One Edge Decreased/ One Component Increased



The previous procedure requires monitor of every edge presented in the trees. When one would like to lower the number of nodes to monitor or if there is a further constraint on nodes that could be sampled, we eventually arrive at what is called the representative CuSum.

4.3.1 Proposed Procedure

$$T_{R-D} = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}_R} \min_{\theta_{rs}} \left\{ \sum_{i=k}^n \log \frac{1}{1 + \tanh \theta_{rs} x_r^{(i)} x_s^{(i)}} - \log(\cosh \theta_{rs}) C \right\} \geq 0 \right\}. \quad (4.23)$$

Minimization domain of θ_{rs} depends on sign of $\prod_{\mu \in \text{path}(r,s)} \theta_{\mu}$ and the length of $\text{path}(r, s)$; if $\prod_{\mu \in \text{path}(r,s)} \theta_{\mu} < 0$, minimization is taken over $-\infty < -\theta_+ \leq \theta_{rs} \leq -\tanh^{-1} \left(\tanh(\theta_-)^{\text{len}(\text{path}(r,s))} \right) < 0$; otherwise, $0 < \tanh^{-1} \left(\tanh(\theta_-)^{\text{len}(\text{path}(r,s))} \right) \leq \theta_{rs} \leq \theta_+ < \infty$.

Similarly as in subsection 4.2.2, the inner minimization yields:

$$\text{If } \prod_{\mu \in \text{path}(r,s)} \theta_{\mu} < 0,$$

When $C > n - k + 1$, minimum attains at $-\theta_+$; when $C \leq n - k + 1$, minimum attains at $-\tanh^{-1} \left(\tanh(\theta_-)^{\text{len}(\text{path}(r,s))} \right)$.

$$\text{If } \prod_{\mu \in \text{path}(r,s)} \theta_{\mu} \geq 0,$$

When $C > n - k + 1$, minimum attains at θ_+ ; when $C \leq n - k + 1$, minimum attains at $\tanh^{-1} \left(\tanh(\theta_-)^{\text{len}(\text{path}(r,s))} \right)$.

Therefore, the equivalent expression as suggested by Mei:

$$T_{R-D} = \inf\{n \geq 1 : \max_{n-C+1 \leq k \leq n} \max_{(r,s) \in \mathcal{E}_R} \sum_{i=k}^n \log\left(\frac{1}{1 + \tanh(\theta_-)^{\text{len}(\text{path}(r,s))} x_r^{(i)} x_s^{(i)}}\right)\} \quad (4.24)$$

$$\geq \log \left(\cosh(\text{sgn}(\theta_{rs}) \tanh^{-1} \left(\tanh(\theta_-)^{\text{len}(\text{path}(r,s))} \right)) \right) C \text{ or} \quad (4.25)$$

$$\max_{(r,s) \in \mathcal{E}} W_{n-C;(r,s)} + \sum_{n-C+1}^n \log\left(\frac{1}{1 + \tanh(\text{sgn}(\theta_{rs})\theta_+) x_r^{(i)} x_s^{(i)}}\right) \quad (4.26)$$

$$\geq \log \left(\cosh(\text{sgn}(\theta_{rs})\theta_+) \right) C \}. \quad (4.27)$$

where $W_{k;(r,s)} = \left(W_{k-1;(r,s)} + \log\left(\frac{1}{1 + \tanh(\text{sgn}(\theta_{rs})\theta_+) x_r^{(i)} x_s^{(i)}}\right) \right)^+$.

4.3.2 Performance of Algorithm

Same as in the previous section, the proposed algorithm is evaluated under the reverse of Lorden's criteria by studying the algorithm's worst average detection delay and average run length to false alarm.

Theorem 4.3.1. *Upper bound to worst average detection delay and lower bound to average run length to false alarm:*

$$WADD(T_{R-D}) \leq C(1 + o(1)) \text{ as } C \rightarrow \infty. \quad (4.28)$$





$$\mathbf{ARL2FA}(T_{R-D}) \geq \frac{2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \text{path}(r,s)} \tanh \theta_\mu)|))}}{|\mathcal{E}_R|} \quad (4.29)$$

$$= \frac{\cosh(\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \text{path}(r,s)} \tanh \theta_\mu)|)}{|\mathcal{E}_R|}. \quad (4.30)$$

Proof. Proof technique follows exactly that of theorem4.2.1. □

Theorem 4.3.2. *With threshold $C = \alpha$, achievability result is established:*

$$\mathbf{WADD}(T_{R-D}) \leq \alpha; \mathbf{ARL2FA}(T_{R-D}) \geq 2^{\log(\cosh(\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \text{path}(r,s)} \tanh \theta_\mu)|))\alpha} \text{ as } \alpha \rightarrow \infty. \quad (4.31)$$

Therefore, T_{R-D} is not asymptotically first-order optimal under reverse Lorden's criteria.

Proof. The first part of the theorem follows directly from theorem4.3.1. □

The algorithm has inferior statistical guarantee compared to that of the previous section with the benefit of monitoring less nodes. Comparison of statistical performance:

$$D(f_1||f_0) \geq \log(\cosh(\min_{(r,s) \in \mathcal{E}} |\theta_{rs}|)) \geq \log(\cosh(\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \text{path}(r,s)} \tanh \theta_\mu)|))). \quad (4.32)$$

The procedure is of course not instance optimal, and also no longer optimal in the sense of definition2.1.10.



4.3.3 Choosing the Representatives

Previously we show that instead of monitoring all the edges presented in the pre-change tree if we only monitor a pre-chosen set of nodes, there's a trade-off between statistical guarantee and computational complexity. A natural question to ask: How to construct a good set of tuples \mathcal{E}_R ? The construction should base on three principles:

1. Every edge has to be covered (i.e. \mathcal{E}_R is chosen such that $\forall e \in \mathcal{E}, \exists (a, b) \in \mathcal{E}_R$ with $e \in \mathbf{path}(a, b)$). This rule is required for the proof of **WADD**(T_{R-D}).
2. Construct \mathcal{E}_R such that $\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \mathbf{path}(r,s)} \tanh \theta_\mu)|$ is as large as possible.
3. Monitor as less tuples as possible (i.e. small cardinality for \mathcal{E}_R).

In order to monitor any possible deletion in the tree, bullet one is necessary for the design of \mathcal{E}_R . The condition from bullet two makes sure that the performance of the algorithm is the best under the worst possible location of the deletion. Bullet three calls for the smallest number of monitors. Note that bullets two and three representing respectively statistical performance and computational complexity can be traded off.

Here we provide a heuristic satisfying bullet one, optimizing over the condition from bullet two with a decent constraint from bullet three. Specifically, our goal is to monitor a number of tuples equal to the number of leaves in the tree which satisfies bullet one and has the maximal possible $\min_{(r,s) \in \mathcal{E}_R} |\tanh^{-1}(\prod_{\mu \in \mathbf{path}(r,s)} \tanh \theta_\mu)|$. Since at least $\lceil \frac{\text{num. of leaves}}{2} \rceil$ is required to cover every edge, our algorithm is decent in the sense that only twice the least number of monitors is required.

Our algorithm first finds a node central to tree \mathcal{T} , call it l , and subsequently monitors

every tuple of the form (l, u) with $u \in \text{leaves}(\mathcal{T})$. The rule certifies that the cardinality of \mathcal{E}_R will equal the number of leaves in the tree. To optimize over the condition from bullet two, the central node l should be judiciously chosen. The node l is chosen via solving:

$$\arg \max_{l \in \mathcal{T}} \min_{u \in \mathcal{T}} \left| \prod_{\mu \in \text{path}(l, u)} \tanh(\theta_\mu) \right|. \quad (4.33)$$

We have already seen the same exact problem in subsection 3.4.3. There we provide a naive algorithm to solve the problem:

1. $\forall i \in \mathcal{T}$, recursively solve $\min_{u \in \mathcal{T}} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)|$.

$$\min_{u \in \mathcal{T}} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)| = \min_{u \in \text{leaves}(\mathcal{T})} \prod_{\mu \in \text{path}(i, u)} |\tanh(\theta_\mu)| \quad (4.34)$$

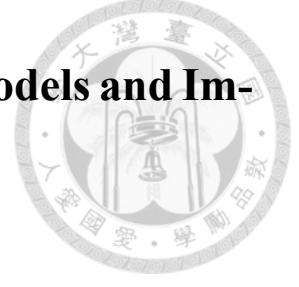
$$= \min_{a \in N(i)} \left\{ \tanh(\theta_{ia}) \min_{u \in \text{leaves}(\mathcal{T}_{ai})} \prod_{\mu \in \text{path}(a, u)} |\tanh(\theta_\mu)| \right\}. \quad (4.35)$$

where $N(i)$ denotes the set of neighboring nodes of i , \mathcal{T}_{ai} is sub-tree of \mathcal{T} constructed from breaking edge θ_{ai} and retaining the component with node a .

2. Take max over $i \in \mathcal{T}$.

Of course the pre-change weights are not known in advance; therefore, in practice, the algorithm is run with constant θ_μ for all μ .

4.4 One Edge Decreased in General Ising Models and Implications



In this section, we first extend the problem setting of section 4.1 to the case where the underlying model is now the general Ising model. Next, we will discuss results from section 4.2 from the perspective of edge deletion in general Ising models. The discussion finally leads to the conclusion that there exists a more general family of Ising model for which the weights are not necessary to achieve good quickest change detection performance. We will give one example beyond Ising forests.

4.4.1 Extension from Section 4.1

The extended setting from section 4.1 of detecting an edge decrease in the general Ising model with the decision-maker knowing the pre-change distribution is also a special case of the quickest change detection with multiple post-change distribution problem. Generalized CuSum can be shown to be asymptotically optimal by adopting analysis from subsection 4.1.3. From an algorithmic design perspective, however, computational burden innate to generalized CuSum arises due to the calculation of likelihood ratios.

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{Z_0}{Z_1} \exp\{-\theta_{rs} x_r x_s\} = \exp\{-\theta_{rs} x_r x_s\} \frac{A_{rs} + B_{rs}}{e^{-\theta_{rs}} A_{rs} + e^{\theta_{rs}} B_{rs}} \quad (4.36)$$

where $A_{rs} = \sum_{\mathbf{x}; x_r x_s = +1} \exp\{\sum_{v \in \mathcal{E}} \delta_v x_v + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t\}$ and $B_{rs} = \sum_{\mathbf{x}; x_r x_s = -1} \exp\{\sum_{v \in \mathcal{E}} \delta_v x_v + \sum_{(u,t) \in \mathcal{E}} \theta_{ut} x_u x_t\}$.

Generalized CuSum yields:

$$W_{rs}^{(1)} = 0; W_{rs}^{(n+1)} = (W_{rs}^{(n)} - \theta_{rs} x_r^{(n+1)} x_s^{(n+1)} + \log(\frac{A_{rs} + B_{rs}}{e^{-\theta_{rs}} A_{rs} + e^{\theta_{rs}} B_{rs}}), 0)^+ \forall (r, s) \in \mathcal{E}. \quad (4.37)$$

$$T_{G,rs} = \inf\{n \geq 1 : W_{rs}^{(n)} \geq C\}; T_G = \min_{(r,s) \in \mathcal{E}} \{T_{G,rs}\}. \quad (4.38)$$

4.4.2 Discussion of Extension from Section 4.2

In section 4.2, it's shown that the edge deletion problem in Ising model on a forest is closely related to the pre-change unknown Bernoulli quickest change detection problem (i.e. shifting from $\mathbf{Ber}(\mu)$ with $\mu \neq \frac{1}{2}$ to $\mathbf{Ber}(\frac{1}{2})$). It turns out for general Ising models, the pattern remains.

$$\frac{f_1(x)}{f_0(x)} = e^{-\theta_{rs} x_r x_s} \frac{Z_0}{Z_1} = \frac{e^{-\theta_{rs}} A_{rs} + e^{\theta_{rs}} B_{rs}}{e^{\theta_{rs} x_r x_s} (A_{rs} + B_{rs})}. \quad (4.39)$$

where $A_{rs} = \sum_{\{\mathbf{x}: x_r x_s = +1\}} e^{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E} - \theta_{ut} x_u x_t}$ and

$$B_{rs} = \sum_{\{\mathbf{x}: x_r x_s = -1\}} e^{\sum_{s \in \mathcal{V}} \delta_s x_s + \sum_{(u,t) \in \mathcal{E} - \theta_{ut} x_u x_t}.$$

The likelihood ratio hints that the problem is equivalent to the Bernoulli mean shift problem $\mathbf{Ber}(\frac{e^{-\theta_{rs}} A_{rs}}{e^{-\theta_{rs}} A_{rs} + e^{\theta_{rs}} B_{rs}})$ to $\mathbf{Ber}(\frac{A_{rs}}{A_{rs} + B_{rs}})$. When the pre/post-change weights are not known to the decision-maker, $\theta_{rs}, A_{rs}, B_{rs}$ are all not known, and thus the problem is extremely difficult. However, when restricted to Ising models on a forest, as in section 3.2, it can be deduced that we always have $A_{rs} = B_{rs}$ by algebra. Consequently, even without knowing the pre/post-change weights, we can always transform the problem into a pre-

change known, post-change unknown Bernoulli mean shift problem.

A natural question to be further asked: Does there exist a class of distribution broader than Ising models on a forest such that $\mathbf{Ber}(\frac{A_{rs}}{A_{rs}+B_{rs}})$ also reduces to $\mathbf{Ber}(\frac{1}{2})$? If so, all results from section 4.2 directly apply to the new class of distribution. Here we give an example. A broader class of Ising models equipped with the aforementioned property is any Ising model with zero mean-field vector $\delta_s = 0 \ \forall s \in \mathcal{V}$ for which only bridges are allowed to disappear after the change. Edge deletion in Ising forests is clearly just a special case of the broader setting.

4.5 Some Lemmas

Define the sequential hypothesis test for pre-change unknown and post-change known distributions:

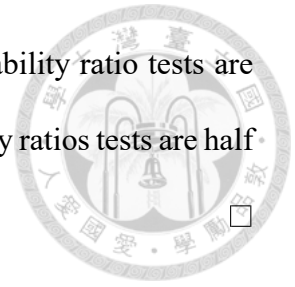
$$\tau_{Mei}(C) = \inf \left\{ n \geq 1 : \sum_{i=1}^n \log \frac{f_\lambda(x^{(i)})}{f_\delta(x^{(i)})} \geq D(f_\lambda \| f_\delta) C \text{ for all } -\infty < \delta_- \leq \delta \leq \delta_+ < \lambda \right\}. \quad (4.40)$$

Lemma 4.5.1 (Mei's Lemma). *When f_λ, f_δ belongs to one-parameter exponential family $f_\gamma(x) = \exp\{\gamma x - b(\gamma)\}$, under f_λ*

$$\mathbb{E}_\lambda[\tau_{Mei}] = C + \sqrt{\frac{b''(\lambda)}{2\pi}} \left(\frac{\lambda - \delta_-}{D(f_\lambda \| f_{\delta_-})} - \frac{\lambda - \delta_+}{D(f_\lambda \| f_{\delta_+})} + o(1) \right) \sqrt{C} \text{ as } C \rightarrow \infty. \quad (4.41)$$

Proof. The lemma is from theorem 2.1 of [14]. Here we provide the high-level idea of the proof. The proof is based on the critical fact that τ_{Mei} reduces to sequential probability ratio test with pre-change parameter equal δ_- when $C > n$ and δ_+ when $C \leq n$. Further, it's argued that the conditioned expected stopping time before and after $n = C$ is some

constant multiple of each other. Using the fact that sequential probability ratio tests are asymptotically normal and after conditioning the sequential probability ratios tests are half-normal distributed, the proof can be concluded. \square





Chapter 5 Conclusion

5.1 Summary

In this thesis, we first posed the structural changing graphical model quickest change detection problem and thereafter studied two types of structural change, edge increase and decrease, on Ising models. We showed that generalized CuSum with the aid of only pre-change structural information achieves the optimal worst average detection delay and average run length to false alarm trade-off for the problem of detecting multiple increased edges in Ising models on a forest. We then turn to exploit the opportunity of correlation propagation unique to correlated models and show how to lower significantly the amount of data sampled at the cost of moderate statistical performance. Later we extended our study to one edge increase in general Ising models and demonstrated its connection to the Bernoulli change point detection problem, thereby explaining why instance optimality can be achieved for a certain family of Ising models without knowing the pre-change weights. For the edge decreased setting, we, unlike the edge increased setting, when equipped with only structure-related information, were only able to come up with a sub-optimal detection procedure. Next, we again proposed a node sampling scheme to bring down the computational burden for detection by leveraging the spatial correlations. Finally, our study of the one edge decrease in general Ising models problem allows us to directly extend results

for Ising models on a forest to a broader class of distribution.



5.2 Future Work

Our original goal for the thesis is to study whether knowledge of the pre-change structure suffices for good detection of structural changing graphical models. Results established in this thesis prove the conjecture positive only when restricting to specific sub-classes. A possible future direction is to generalize the result to broader classes of graphical models. Towards this end, we speculate going beyond variants of likelihood ratio statistics and using statistics from the graphical model learning literature would be helpful.

For both the increase and decrease edge problem setting, we have posed the problem of monitoring only a few judiciously selected nodes while maintaining decent statistical performance. The problem can also be viewed as a passive sampling node selection problem. Our results show that if the decision-maker is bound to sample the same nodes during the entire detection time duration, we can only crave for sub-optimal performance. A natural future direction is to allow the decision-maker to actively sample nodes from the network and study its performance. Under the active sampling setting, we could even give additional constraints such as restricting the nodes to be sampled each time-stamp has to be a neighbor of the node sampled in the previous time-stamp.

As mentioned, the propagation of correlation serves as an opportunity for the quickest detection task. We showed in the thesis, how to take advantage of the property for graphical models with fast correlation decay, models which are known to be simpler to learn. We speculate that for the quickest change detection problem, on the other hand, the

presence of long-range correlation is beneficial. Intuitively, slower decay of correlation should give the chance of detecting the change promptly even when the location of the change is erroneously guessed. We, however, expect that detection procedures are harder to design for such models. It would be an interesting future work to make the intuition rigorous.

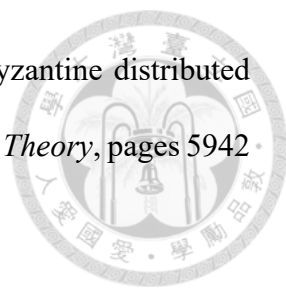
Finally, a future direction is to close the gap between the achievability and converse result for the one edge decreased in Ising model on a forest problem. Since the problem is fundamentally a quickest change detection problem with composite pre-change distribution and a finite number of possible post-change distributions, a conclusive result would be a novel contribution to the quickest detection research area.

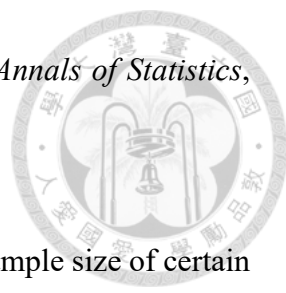




References

- [1] B. L. Bars, P. Humbert, A. Kalogeratos, and N. Vayatis. Learning the piece-wise constant graph structure of a varying ising model. In *International Conference of Machine Learning*. 2020.
- [2] J. Bento and A. Montanari. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*, pages 1303–1311. 2009.
- [3] G. Bresler and M. Karzand. Learning a tree-structured ising model in order to make predictions. In *Annals of Statistics*, pages 713–737. 2020.
- [4] C. Daskalakis, N. Dikkala, and G. Kamath. Testing ising models. In *IEEE Transaction of Information Theory*, pages 6829–6852. 2019.
- [5] F. Fazayeli and A. Banerjee. Generalized direct change estimation in ising model structure. In *International Conference of Machine Learning*, pages 2281–2290. 2016.
- [6] G. Fellouris, E. Bayraktar, and L. Lai. Efficient byzantine sequential change detection. In *IEEE Transactions on Information Theory*, page 3346–3360. 2018.
- [7] A. Gangrade, B. Nazer, and V. Saligrama. Limits on testing structural changes in ising models. In *Advances in Neural Information Processing Systems*. 2020.

- 
- [8] Y. Huang, Y. Huang, and S. Lin. Asymptotic optimality in byzantine distributed quickest change detection. In *IEEE Transactions on Information Theory*, pages 5942 – 5962. 2021.
- [9] A. Katiyar, V. Shah, and C. Caramanis. Robust estimation of tree structured ising models. In *arXiv preprint*. 2020.
- [10] D. Koller and N. Friedman. In *Probabilistic Graphical Models: Principles and Techniques*. 2009.
- [11] K. Liu, R. Zhang, and Y. Mei. Scalable sum-shrinkage schemes for distributed monitoring large- scale data streams. In *Statistica Sinica*, pages 1–22. 2019.
- [12] G. Lorden. Procedures for reacting to a change in distribution. In *The Annals of Mathematical Statistics*, pages 1897–1908. 1971.
- [13] G. Lorden and M. Pollak. Sequential change-point detection procedures that are nearly optimal and computationally simple. In *Sequential Analysis*, pages 476–512. 2008.
- [14] Y. Mei. Sequential change-point detection when unknown parameters are present in the pre-change distribution. In *Annals of Statistics*, pages 92–122. 2006.
- [15] Y. Mei. Efficient scalable schemes for monitoring a large number of data streams. In *Biometrika*, pages 419–433. 2010.
- [16] G. Moustakides. Optimal stopping times for detecting changes in distributions. In *Annals of Statistics*, pages 1379–1387. 1986.
- [17] E. Page. Continuous inspection schemes. In *Biometrika*, pages 100–115. 1954.

- 
- [18] M. Pollak. Optimal detection of a change in distribution. In *Annals of Statistics*, pages 206–227. 1985.
- [19] M. Pollak and D. Siegmund. Approximations to the expected sample size of certain sequential tests. In *Annals of Statistics*, pages 1267–1282. 1975.
- [20] A. Shiryaev. On optimum methods in quickest detection problems. In *Theory of Probability and its Applications*, pages 22–46. 1963.
- [21] D. Siegmund and E. Venkatraman. Using the generalized likelihood ratio statistics for sequential detection of a change-point. In *Annals of Statistics*, pages 255–271. 1995.
- [22] N. Suh, R. Zhang, and Y. Mei. Adaptive online monitoring of the ising model. In *Annual Allerton Conference on Communication, Control, and Computing*. 2019.
- [23] J. Unnikrishnan, V. Veeravalli, and S. Meyn. Minimax robust quickest change detection. In *IEEE Transaction of Information Theory*, pages 1604–1614. 2011.
- [24] V. V. Veeravalli and T. Banerjee. Quickest change detection. In *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*, pages 209–256. 2013.
- [25] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. In *Foundations and Trends in Machine Learning*, pages 1–305. 2008.
- [26] L. Xie, Y. Xie, and G. V. Moustakides. Sequential subspace change point detection. In *Sequential Analysis*, pages 307–335. 2020.

- [27] L. Xie, S. Zou, Y. Xie, and V. Veeravalli. Sequential (quickest) change detection: Classical results and new directions. In *IEEE Journal on Selected Areas in Information Theory*. 2021.

