© 2019 by Yanglei Song. All rights reserved.

SOME TOPICS IN SEQUENTIAL ANALYSIS

BY

YANGLEI SONG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Assistant Professor Georgios Fellouris, Chair and Director of Research
Professor Jeffrey Douglas
Professor Emeritus Adam Martinsek
Professor Venugopal Veeravalli

Abstract

Sequential analysis refers to the statistical theory and methods that can be applied to situations where the sample size is not fixed in advance. Instead, the data are collected sequentially over time, and the sampling is stopped according to a pre-specified stopping rule as soon as the accumulated information is deemed sufficient. The goal of this adaptive approach is to reach a reliable decision as soon as possible. This dissertation investigates two problems in sequential analysis.

In the first problem, assuming that data are collected sequentially from independent streams, we consider the simultaneous testing of multiple hypotheses. We start with the class of procedures that control the classical familywise error probabilities of both type I and type II under two general setups: when the number of signals (correct alternatives) is known in advance, and when we only have a lower and an upper bound for it. Then we continue to study two generalized error metrics: under the first one, the probability of at least k mistakes, of any kind, is controlled; under the second, the probabilities of at least k_1 false positives and at least k_2 false negatives are simultaneously controlled. For each formulation, the optimal expected sample size is characterized, to a first-order asymptotic approximation as the error probabilities vanish, and a novel multiple testing procedure is proposed and shown to be asymptotically efficient under every signal configuration.

In the second problem, we propose a generalization of the Bayesian sequential change detection problem, where the change is a latent event that should be not only detected but also accelerated. It is assumed that the sequentially collected observations are responses to treatments selected in real time. The assigned treatments not only determine the distribution of responses before and after the change, but also influence when the change happens. The problem is to find a treatment assignment rule and a stopping rule to minimize the average total number of observations subject to a bound on the false-detection probability. We propose an intuitive solution, which is easy to implement and achieves for a large class of change-point models the optimal performance up to a first-order asymptotic approximation. A simulation study suggests the almost exact optimality of the proposed scheme under a Markovian change-point model.

To my wife, Qian Lu, and daughter, Monica Xuan Song.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Professor Georgios Fellouris for his continuous support and tremendous guidance. I was introduced to the area of sequential analysis by him, on which I focus in my Ph.D. study. Over these years, his insightful discussions and suggestions have benefited not only my research but also my teaching, career, and life. It has been a great pleasure to work with him.

Besides my advisor, I would like to thank my other thesis committee: Professor Jeffrey Douglas, Professor Adam Martinsek and Professor Venugopal Veeravalli, for their time, support and helpful comments. In addition, I would like to thank Professor Xiaohui Chen for his advice on additional research topics. My thanks also go to all the faculty, staff members and fellow students in the Statistics department, who are an integral part of my life in graduate school.

Finally, I would like to thank my family, particularly my wife. Without her unconditional support, this dissertation would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Two classical problems in sequential analysis are sequential hypothesis testing, initiated by Wald's seminal paper [75], and quickest change-point detection, pioneered by Shwehart [61] and Page [51]. We refer interested readers to [35, 71] for an extensive review on the theory, methodology and diverse applications of sequential analysis. In this dissertation, we study extensions of the classical problems, which are briefly discussed below and developed in detail in the following chapters.

1.1 Multiple testing with sequential data

When testing simultaneously *multiple* hypotheses with data collected from a different stream for each hypothesis, there are two natural generalizations of Wald's sequential framework [75]. In the first one, sampling can be terminated earlier in some data streams [3, 7, 45]. In the second, which is the focus of Chpater 2 and 3, sampling is terminated at the same time in all streams [17, 18]. The latter setup is motivated by applications such as multichannel signal detection [73], multiple access wireless network [57] and multisensor surveillance systems [26], where a centralized decision maker needs to make a decision regarding the presence or absence of signal, e.g., an intruder, in multiple channels/areas monitored by a number of sensors. This framework is also motivated by online surveys and crowdsourcing tasks [33], where the goal is to find "correct" answers to a fixed number of questions, e.g., regarding some product or service, by asking the smallest necessary number of people.

Specifically, we consider J data streams, each associated with a hypothesis testing problem. At any time prior to stopping, we collect one observation from each stream, and we decide whether to continue or to stop the sampling process based on the current and past observations; in the latter case, we need to solve all J problems based on the information prior to stopping (see Figure 1.1).

In Chapter 2, we consider the class of procedures that control the classical familywise error probabilities of both type I and type II below given, user-specified levels, under two general setups: when the number of signals (correct alternatives) is known in advance, and when we only have a lower and an upper bound for

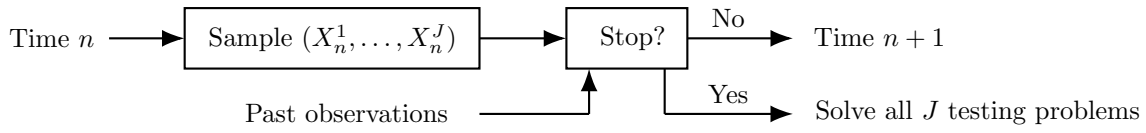


Figure 1.1: J data streams. X_n^j is the observation collected from j-th stream at time n.

it. In Chapter 3, we consider two generalized error metrics: *i*) the probability of at least k mistakes of any kind; *ii*) the probabilities of at least k_1 false positives and at least k_2 false negatives.

For each above formulation, under the independent streams assumption, we 1) characterize the optimal expected sample size asymptotically as the error probabilities vanish, 2) propose a novel, feasible procedure with non-asymptotic error control, 3) establish its asymptotic efficiency, and 4) quantify the gains of sequential sampling over fixed-sample schemes.

1.2 Change detection with experimental design

Quickest change detection (QCD), the problem of detecting a change in the statistical properties of streaming data, arises in applications such as quality monitoring, threat detection, and epidemic control. In the literature, there are two main formulations: *i*) the mechanism that triggers the change is unknown; *ii*) the change-point follows some prior distribution, and is not affected by observations. Thus, it is neither permissible nor relevant to influence the change-point, which restricts the applicability of QCD in some situations. We are in particular motivated by applications in intelligent tutoring systems, and we propose a new paradigm where the change should be not only detected, but also accelerated.

Specifically, in Chapter 4, we consider a latent binary process $\{L_t\}$, whose value transits to one at some unknown change-point (see Figure 1.2). At each time t, we select a treatment X_t among a number of options and observe a response Y_t whose distribution depends on X_t and the latent status L_t. Then, based on the collected responses up to this time, we decide whether to stop and declare that a change has occurred, or to continue the process, in which case we have to decide the treatment for time $t+1$. We assume that the change is irreversible, and the probability of change at current time is a function of past treatments. Our goal is to find a treatment assignment rule and a stopping rule to minimize the average number of observations subject to a bound on the false detection probability. For a class of change-point models, we obtain the optimal solution using dynamic programming, which however is not always computationally feasible and can only be obtained numerically. Thus, we propose a novel procedure whose structure is explicit and whose thresholds are specified via minimizing an upper bound on the sampling cost. In addition, we establish its asymptotic efficiency under certain conditions.

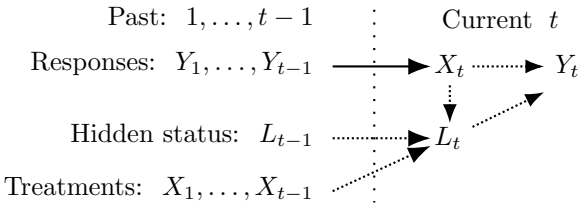


Figure 1.2: An assignment rule selects treatment X_t based on past responses. If $L_{t-1} = 0$, the probability that $L_t = 1$ depends on the treatments up to time t.

Chapter 2

Sequential multiple testing with prior information

2.1 Introduction

[1] Multiple testing, that is the simultaneous consideration of K hypothesis testing problems, H_0^k versus H_1^k, $1 \leq k \leq K$, is one of the oldest, yet still very active areas of statistical research. The vast majority of work in this area assumes a fixed set of observations and focuses on testing procedures that control the familywise type I error (i.e., at least one false positive), as in [28, 29, 46], or less stringent metrics of this error, as in [8] and [36].

The multiple testing problem has been less studied under the assumption that observations are acquired sequentially, in which case the sample size is random. The sequential setup is relevant in many applications, such as multichannel signal detection [21, 47], outlier detection [40], clinical trials with multiple end-points [4], ultra high throughput mRNA sequencing data [6], in which it is vital to make a quick decision in real time, using the smallest possible number of observations.

Bartroff and Lai [5] were the first to propose a sequential test that controls the familywise error of type I. De and Baron [17, 18] and Bartroff and Song [7] proposed universal sequential procedures that control simultaneously the familywise errors of both type I *and* type II, a feature that is possible due to the sequential nature of sampling. The proposed sequential procedures in these works were shown through simulation studies to offer substantial savings in the average sample size in comparison to the corresponding fixed-sample size tests.

A very relevant problem to multiple testing is the classification problem, in which there are M hypotheses, $H_1, \ldots, H_M$, and the goal is to select the correct one among them. The classification problem has been studied extensively in the literature of sequential analysis, see e.g. [1, 21, 22, 44, 64, 72], generalizing the seminal work of Wald [75] on binary testing ($M = 2$). Dragalin et al. [22] considered the multiple testing problem as a special case of the classification problem under the assumption of a *single signal* in K independent streams, and focused on procedures that control the probability of erroneously claiming the

[1] This chapter is based on my publication [66].

signal to be in stream i for every $1 \leq i \leq M = K$. In this framework, they proposed an asymptotically optimal sequential test as all these error probabilities go to 0. The same approach of treating the multiple testing problem as a classification problem has been taken by Li et al. [40] under the assumption of an upper bound on the number of signals in the K independent streams, and a *single control* on the maximal mis-classification probability.

We should stress that interpreting multiple testing as a classification problem does not generally lead to feasible procedures. Consider, for example, the case of no prior information, which is the default assumption in the multiple testing literature. Then, multiple testing becomes a classification problem with $M = 2^K$ categories and a brute-force implementation of existing classification procedures becomes infeasible even for moderate values of K, as the number of statistics that need to be computed sequentially grows exponentially with K. Independently of feasibility considerations, to the best of our knowledge there is no optimality theory regarding the expected sample size that can be achieved by multiple testing procedures, with or without prior information, that control the familywise errors of both type I and type II. Filling this gap was one of the motivations of this Chapter.

The main contributions of the current Chapter are the following: first of all, assuming that the data streams that correspond to the various hypotheses are independent, we propose feasible procedures that control the familywise errors of both type I and type II below arbitrary, user-specified levels. We do so under two general setups regarding prior information; when the true number of signals is known in advance, and when there is only a lower and an upper bound for it. The former setup includes the case of a single signal considered in Dragalin et al. [21, 22], whereas the latter includes the case of no prior information, which is the underlying assumption in Bartroff and Song [7], De and Baron [17, 18]. While we provide universal threshold values that guarantee the desired error control in the spirit of the above works, we also propose a Monte Carlo simulation method based on importance sampling for the efficient calculation of non-conservative thresholds in practice, even for very small error probabilities. More importantly, in the case of independent and identically distributed (i.i.d.) observations in each stream, we show that the proposed multiple testing procedures attain the optimal expected sample size, for *any* possible signal configuration, to a first-order asymptotic approximation as the two error probabilities go to zero in an *arbitrary* way. Our asymptotic results also provide insights about the effect of prior information on the number of signals, which are corroborated by a simulation study.

The remainder of the Chapter is organized as follows. In Section 2.2 we formulate the problem mathematically. In Section 2.3 we present the proposed procedures and show how they can be designed to guarantee the desired error control. In Section 2.4 we propose an efficient Monte Carlo simulation method for the

determination of non-conservative critical values in practice. In Section 2.5 we establish the asymptotic optimality of the proposed procedures in the i.i.d. setup. In Section 2.6 we illustrate our asymptotic results with a simulation study. In Section 2.7 we conclude and discuss potential generalizations of our work. Finally, we present two useful lemmas for our proofs in Section 2.8.

2.2 Problem formulation

Consider K independent streams of observations, $X^k := \{X_n^k : n \in \mathbb{N}\}$, $k \in [K]$, where $[K] := \{1, \ldots, K\}$ and $\mathbb{N} := \{1, 2, \ldots\}$. For each $k \in [K]$, let P^k be the distribution of X^k, for which we consider two simple hypotheses,

$$H_0^k : \mathsf{P}^k = \mathsf{P}_0^k \text{ versus } H_1^k : \mathsf{P}^k = \mathsf{P}_1^k,$$

where P_0^k and P_1^k are distinct probability measures on the canonical space of X^k. We will say that there is "noise" in the k^{th} stream under P_0^k and "signal" under P_1^k. Our goal is to simultaneously test these K hypotheses when data from all streams become available sequentially and we want to make a decision as soon as possible.

Let $\mathcal{F}_n$ be the σ-field generated by all streams up to time n, i.e., $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, where $X_n = (X_n^1, \ldots, X_n^K)$. We define a *sequential* test for the multiple testing problem of interest to be a pair (T, d) that consists of an $\{\mathcal{F}_n\}$-stopping time, T, at which we stop sampling in all streams, and an $\mathcal{F}_T$-measurable decision rule, $d = (d^1, \ldots, d^K)$, each component of which takes values in $\{0, 1\}$. The interpretation is that we declare upon stopping that there is signal (resp. noise) in the k^{th} stream when $d^k = 1$ (resp. $d^k = 0$). With an abuse of notation, we will also use d to denote the subset of streams in which we declare that signal is present, i.e., $\{k \in [K] : d^k = 1\}$.

For any subset $\mathcal{A} \subset [K]$ we define the probability measure

$$\mathsf{P}_{\mathcal{A}} := \bigotimes_{k=1}^{K} \mathsf{P}^k; \qquad \mathsf{P}^k = \begin{cases} \mathsf{P}_0^k, & \text{if } k \notin \mathcal{A} \\ \mathsf{P}_1^k, & \text{if } k \in \mathcal{A} \end{cases},$$

such that the distribution of $\{X_n, n \in \mathbb{N}\}$ is $\mathsf{P}_{\mathcal{A}}$ when $\mathcal{A}$ is the true subset of signals, and for an arbitrary

sequential test (T, d) we set:

$$\{\mathcal{A} \lesssim d\} := \{(d \setminus \mathcal{A}) \neq \emptyset\} = \bigcup_{j \notin \mathcal{A}} \{d^j = 1\},$$
$$\{d \lesssim \mathcal{A}\} := \{(\mathcal{A} \setminus d) \neq \emptyset\} = \bigcup_{k \in \mathcal{A}} \{d^k = 0\}.$$

Then, $\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d)$ is the probability of at least one false positive (*familywise type I error*) and $\mathsf{P}_{\mathcal{A}}(d \lesssim \mathcal{A})$ the probability of at least one false negative (*familywise type II error*) of (T, d) *when the true subset of signals is* $\mathcal{A}$.

In this Chapter we are interested in sequential tests that control these probabilities below user-specified levels α and β respectively, where $\alpha, \beta \in (0, 1)$, for any possible subset of signals. In order to be able to incorporate prior information, we assume that the true subset of signals is known to belong to a class $\mathcal{P}$ of subsets of $[K]$, not necessarily equal to the powerset, and we focus on sequential tests in the class

$$\Delta_{\alpha,\beta}(\mathcal{P}) := \{(T, d) : \mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d) \leq \alpha \text{ and } \mathsf{P}_{\mathcal{A}}(d \lesssim \mathcal{A}) \leq \beta \text{ for every } \mathcal{A} \in \mathcal{P}\}.$$

We consider, in particular, two general cases for class $\mathcal{P}$. In the first one, it is known that there are exactly m signals in the K streams, where $1 \leq m \leq K - 1$. In the second, it is known that there are at least ℓ and at most u signals, where $0 \leq \ell < u \leq K$. In the former case we write $\mathcal{P} = \mathcal{P}_m$ and in the latter $\mathcal{P} = \mathcal{P}_{\ell,u}$, where

$$\mathcal{P}_m := \{\mathcal{A} \subset [K] : |\mathcal{A}| = m\}, \quad \mathcal{P}_{\ell,u} := \{\mathcal{A} \subset [K] : \ell \leq |\mathcal{A}| \leq u\}.$$

When $\ell = 0$ and $u = K$, the class $\mathcal{P}_{\ell,u}$ is the powerset of $[K]$, which corresponds to the case of no prior information regarding the multiple testing problem.

Our main focus is on multiple testing procedures that not only belong to $\Delta_{\alpha,\beta}(\mathcal{P})$ for a given class $\mathcal{P}$, but also achieve the minimum possible expected sample size, under each possible signal configuration, for small error probabilities. To be more specific, let $\mathcal{P}$ be a given class of subsets and let (T^*, d^*) be a sequential test that can designed to belong to $\Delta_{\alpha,\beta}(\mathcal{P})$ for any given $\alpha, \beta \in (0, 1)$. We say that (T^*, d^*) is *asymptotically optimal with respect to class* $\mathcal{P}$, if for every $\mathcal{A} \in \mathcal{P}$ we have as $\alpha, \beta \to 0$

$$\mathsf{E}_{\mathcal{A}}[T^*] \sim \inf_{(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P})} \mathsf{E}_{\mathcal{A}}[T],$$

where $\mathsf{E}_{\mathcal{A}}$ refers to expectation under $\mathsf{P}_{\mathcal{A}}$ and $x \sim y$ means that $x/y \to 1$. The ultimate goal of this Chapter is to propose feasible sequential tests that are asymptotically optimal with respect to classes of the form $\mathcal{P}_m$

and $\mathcal{P}_{\ell,u}$.

2.2.1 Assumptions and notations

Before we continue with the presentation and analysis of the proposed multiple testing procedures, we will introduce some additional notation, and impose some minimal conditions on the distributions in each stream, which we will assume to hold throughout the Chapter.

First of all, for each stream $k \in [K]$ and time $n \in \mathbb{N}$ we assume that the probability measures P_0^k and P_1^k are mutually absolutely continuous when restricted to the σ-algebra $\mathcal{F}_n^k = \sigma(X_1^k, \ldots, X_n^k)$, and we denote by

$$\lambda^k(n) := \log \frac{d\mathsf{P}_1^k}{d\mathsf{P}_0^k}(\mathcal{F}_n^k) \tag{2.1}$$

the cumulative log-likelihood ratio at time n based on the data in the k^{th} stream. Moreover, we assume that for each stream $k \in [K]$ the probability measures P_0^k and P_1^k are singular on $\mathcal{F}_\infty^k := \sigma(\cup_{n\in\mathbb{N}}\mathcal{F}_n^k)$, which implies that

$$\mathsf{P}_0^k\left(\lim_{n\to\infty} \lambda^k(n) = -\infty\right) = \mathsf{P}_1^k\left(\lim_{n\to\infty} \lambda^k(n) = \infty\right) = 1. \tag{2.2}$$

Intuitively, this means that as observations accumulate, the evidence in favor of the correct hypothesis becomes arbitrarily strong. The latter assumption is necessary in order to design procedures that terminate almost surely under every scenario. *We do not make any other distributional assumption until Section 2.5.*

We use the following notation for the ordered, local, log-likelihood ratio statistics at time n:

$$\lambda^{(1)}(n) \geq \ldots \geq \lambda^{(K)}(n),$$

and we denote by $i_1(n), \ldots, i_K(n)$ the corresponding stream indices, i.e.,

$$\lambda^{(k)}(n) = \lambda^{i_k(n)}(n), \text{ for every } k \in [K].$$

Moreover, for every $n \in \mathbb{N}$ we denote by $p(n)$ the number of positive log-likelihood ratio statistics at time n, i.e.,

$$\lambda^{(1)}(n) \geq \ldots \geq \lambda^{(p(n))}(n) > 0 \geq \lambda^{(p(n)+1)}(n) \geq \ldots \geq \lambda^{(K)}(n).$$

For any two subsets $\mathcal{A}, \mathcal{C} \subset [K]$ we denote by $\lambda^{\mathcal{A},\mathcal{C}}$ the log-likelihood ratio process of $\mathsf{P}_{\mathcal{A}}$ versus $\mathsf{P}_{\mathcal{C}}$, i.e.,

$$\lambda^{\mathcal{A},\mathcal{C}}(n) := \log \frac{d\mathsf{P}_{\mathcal{A}}}{d\mathsf{P}_{\mathcal{C}}}(\mathcal{F}_n) = \sum_{k \in \mathcal{A} \setminus \mathcal{C}} \lambda^k(n) - \sum_{k \in \mathcal{C} \setminus \mathcal{A}} \lambda^k(n), \quad n \in \mathbb{N}. \tag{2.3}$$

Finally, we use $|\cdot|$ to denote set cardinality, for any two real numbers x, y we set $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$, and for any measurable event Γ and random variable Y we use the following notation

$$\mathsf{E}_{\mathcal{A}}[Y; \Gamma] := \int_{\Gamma} Y d\mathsf{P}_{\mathcal{A}}.$$

2.3 Proposed sequential multiple testing procedures

In this section we present the proposed procedures and show how they can be designed in order to guarantee the desired error control.

2.3.1 Known number of signals

In this subsection we consider the setup in which the number of signals is known to be equal to m for some $1 \leq m \leq K-1$, thus, $\mathcal{P} = \mathcal{P}_m$. Without loss of generality, we restrict ourselves to multiple testing procedures (T, d) such that $|d| = m$. Thus, the class of admissible sequential tests takes the form

$$\Delta_{\alpha,\beta}(\mathcal{P}_m) = \{(T, d) : \ \mathsf{P}_{\mathcal{A}}(d \neq \mathcal{A}) \leq \alpha \wedge \beta \text{ for every } \mathcal{A} \in \mathcal{P}_m\},$$

since for any $\mathcal{A} \in \mathcal{P}_m$ and (T, d) such that $|d| = m$ we have

$$\{\mathcal{A} \lesssim d\} \ = \ \{d \lesssim \mathcal{A}\} \ = \ \{d \neq \mathcal{A}\}.$$

In this context, we propose the following sequential scheme: stop as soon as the *gap* between the m-th and $(m+1)$-th ordered log-likelihood ratio statistics becomes larger than some constant $c > 0$, and declare that signal is present in the m streams with the top log-likelihood ratios at the time of stopping. Formally, we propose the following procedure, to which we refer as "gap rule":

$$\begin{aligned} T_G &:= \inf\left\{n \geq 1 : \lambda^{(m)}(n) - \lambda^{(m+1)}(n) \geq c\right\}, \\ d_G &:= \{i_1(T_G), \ldots, i_m(T_G)\}. \end{aligned} \tag{2.4}$$

Here, we suppress the dependence of (T_G, d_G) on m and c to lighten the notation. The next theorem shows how to select threshold c in order to guarantee the desired error control.

Theorem 2.1. *Suppose that assumption* (2.2) *holds. Then, for any* $\mathcal{A} \in \mathcal{P}_m$ *and* $c > 0$ *we have* $\mathsf{P}_A(T_G < \infty) = 1$ *and*

$$\mathsf{P}_{\mathcal{A}}\left(d_G \neq \mathcal{A}\right) \leq m(K-m)e^{-c}. \tag{2.5}$$

Consequently, $(T_G, d_G) \in \Delta_{\alpha,\beta}(\mathcal{P}_m)$ *when threshold* c *is selected as*

$$c = |\log(\alpha \wedge \beta)| + \log(m(K-m)). \tag{2.6}$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_m$ and $c > 0$. We observe that $T_G \leq T_G'$, where

$$\begin{aligned} T_G' &= \inf\left\{n \geq 1 : \lambda^{(m)}(n) - \lambda^{(m+1)}(n) \geq c,\ i_1(n) \in \mathcal{A}, \ldots, i_m(n) \in \mathcal{A}\right\} \\ &= \inf\left\{n \geq 1 : \lambda^{k}(n) - \lambda^{j}(n) \geq c \ \text{ for every } k \in \mathcal{A} \text{ and } j \notin \mathcal{A}\right\}. \end{aligned} \tag{2.7}$$

Due to condition (2.2), it is clear that $\mathsf{P}_{\mathcal{A}}(T_G' < \infty) = 1$, which proves that T_G is also almost surely finite under $\mathsf{P}_{\mathcal{A}}$. We now focus on proving (2.5). The gap rule makes a mistake under $\mathsf{P}_{\mathcal{A}}$ if there exist $k \in \mathcal{A}$ and $j \notin \mathcal{A}$ such that the event $\Gamma_{k,j} = \left\{\lambda^j(T_G) - \lambda^k(T_G) \geq c\right\}$ occurs. In other words,

$$\{d_G \neq \mathcal{A}\} = \bigcup_{k \in \mathcal{A}, j \notin \mathcal{A}} \Gamma_{k,j},$$

and from Boole's inequality we have

$$\mathsf{P}_{\mathcal{A}}(d_G \neq \mathcal{A}) \leq \sum_{k \in \mathcal{A}, j \notin \mathcal{A}} \mathsf{P}_{\mathcal{A}}(\Gamma_{k,j}).$$

Fix $k \in \mathcal{A}, j \notin \mathcal{A}$ and set $\mathcal{C} = \mathcal{A} \cup \{j\} \setminus \{k\}$. Then, from (2.3) we have that $\lambda^{\mathcal{A},\mathcal{C}} = \lambda^k - \lambda^j$ and from Wald's likelihood ratio identity it follows that

$$\begin{aligned} \mathsf{P}_{\mathcal{A}}(\Gamma_{k,j}) &= \mathsf{E}_{\mathcal{C}}\left[\exp\{\lambda^{\mathcal{A},\mathcal{C}}(T_G)\}; \Gamma_{k,j}\right] \\ &= \mathsf{E}_{\mathcal{C}}\left[\exp\{\lambda^{k}(T_G) - \lambda^{j}(T_G)\}; \Gamma_{k,j}\right] \leq e^{-c}, \end{aligned} \tag{2.8}$$

where the last inequality holds because $\lambda^j(T_G) - \lambda^k(T_G) \geq c$ on $\Gamma_{k,j}$. Since $|\mathcal{A}| = m$ and $|\mathcal{A}^c| = K - m$, from the last two inequalities we obtain (2.5), which completes the proof. □

2.3.2 Lower and upper bounds on the number of signals

In this subsection, we consider the setup in which we know that there are at least ℓ and at most u signals for some $0 \leq \ell < u \leq K$, that is, $\mathcal{P} = \mathcal{P}_{\ell,u}$. In order to describe the proposed procedure, it is useful to first introduce the "intersection rule", (T_I, d_I), according to which we stop sampling as soon as *all* log-likelihood ratio statistics are outside the interval $(-a, b)$, and at this time we declare that signal is present (resp. absent) in those streams with positive (resp. negative) log-likelihood ratio, i.e.,

$$\begin{aligned} T_I &:= \inf\left\{n \geq 1 : \lambda^k(n) \notin (-a, b) \text{ for every } k \in [K]\right\}, \\ d_I &:= \{i_1(T_I), \ldots, i_{p(T_I)}(T_I)\}, \end{aligned} \tag{2.9}$$

recalling that $p(n)$ is the number of positive log-likelihood ratios at time n. This procedure was proposed by De and Baron [17], where it was also shown that when the thresholds are selected as

$$a = |\log \beta| + \log K, \quad b = |\log \alpha| + \log K, \tag{2.10}$$

the familywise type-I and type-II error probabilities are bounded by α and β for any possible signal configuration, i.e., $(T_I, d_I) \in \Delta_{\alpha,\beta}(\mathcal{P}_{0,K})$.

A straightforward way to incorporate the prior information of at least ℓ and at most u signals in the intersection rule is to modify the stopping time in (2.9) as follows:

$$\tau_2 := \inf\left\{n \geq 1 : \ell \leq p(n) \leq u \text{ and } \lambda^k(n) \notin (-a, b) \text{ for every } k \in [K]\right\}, \tag{2.11}$$

while keeping the same decision rule as in (2.9). Indeed, stopping according to τ_2 guarantees that the number of null hypotheses rejected upon stopping will be between ℓ and u. However, as we will see in Subsection 2.5.3, this rule will not in general achieve asymptotic optimality in the boundary cases of exactly ℓ and exactly u signals. In order to obtain an asymptotically optimal rule, we need to be able to stop faster when there are exactly ℓ or u signals, which can be achieved by stopping at

$$\begin{aligned} \tau_1 &:= \inf\left\{n \geq 1 : \lambda^{(\ell+1)}(n) \leq -a,\ \lambda^{(\ell)}(n) - \lambda^{(\ell+1)}(n) \geq c\right\}, \\ \text{and} \quad \tau_3 &:= \inf\left\{n \geq 1 : \lambda^{(u)}(n) \geq b,\ \lambda^{(u)}(n) - \lambda^{(u+1)}(n) \geq d\right\}, \end{aligned}$$

respectively. Here, c and d are additional positive thresholds that will be selected, together with a and b, in order to guarantee the desired error control.

We can think of τ_1 as a combination of the intersection rule and the gap rule that corresponds to the case of exactly ℓ signals. Indeed, τ_1 stops when $K-\ell$ log-likelihood ratio statistics are simultaneously below $-a$, but unlike the intersection rule it does not wait for the remaining ℓ statistics to be larger than b; instead, similarly to the gap-rule in (2.4) with $m=\ell$, it requires the gap between the top ℓ and the bottom $K-\ell$ statistics to be larger than c. In a similar way, τ_3 is a combination of the intersection rule and the gap rule that corresponds to the case of exactly u signals.

Based on the above discussion, when we know that there are at least ℓ and at most u signals, we propose the following procedure, to which we refer as "gap-intersection" rule:

$$T_{GI} := \min\{\tau_1, \tau_2, \tau_3\}, \quad d_{GI} := \{i_1(T_{GI}), \dots, i_{p'}(T_{GI})\}, \tag{2.12}$$

where $p' := (p(T_{GI}) \wedge \ell) \vee u$ is a truncated version of the number of positive log-likelihood ratios at T_{GI}, i.e., if $p' = \ell$ when $p(T_{GI}) \leq \ell$, $p' = u$ when $p(T_{GI}) \geq u$ and $p' = p(T_{GI})$ otherwise. In other words, we stop sampling as soon as one of the stopping criterion in τ_1, τ_2 or τ_3 is is satisfied, and we reject upon stopping the null hypotheses in the p' streams with the highest log-likelihood ratio values at time T_{GI}.

As before, we suppress the dependence on ℓ, u and a, b, c, d in order to lighten the notation. Moreover, we set $\lambda^{(0)}(n) = -\infty$ and $\lambda^{(K+1)}(n) = \infty$ for every $n \in \mathbb{N}$, which implies that if $\ell = 0$, then $\tau_1 = \infty$, and if $u = K$, then $\tau_3 = \infty$. When in particular $\ell = 0$ and $u = K$, that is the case of no prior information, $T_{GI} = \tau_2$ and (T_{GI}, d_{GI}) reduces to the intersection rule, (T_I, d_I), defined in (2.9).

The following theorem shows how to select thresholds a, b, c, d in order to guarantee the desired error control for the gap-intersection rule.

Theorem 2.2. *Suppose that assumption* (2.2) *holds. For any subset* $\mathcal{A} \in \mathcal{P}_{\ell,u}$ *and positive thresholds* a, b, c, d, *we have* $\mathsf{P}_A(T_{GI} < \infty) = 1$ *and*

$$\begin{aligned} \mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d_{GI}) &\leq |\mathcal{A}^c| \left(e^{-b} + |\mathcal{A}|\, e^{-c}\right), \\ \mathsf{P}_{\mathcal{A}}(d_{GI} \lesssim \mathcal{A}) &\leq |\mathcal{A}| \left(e^{-a} + |\mathcal{A}^c|\, e^{-d}\right). \end{aligned} \tag{2.13}$$

In particular, $(T_{GI}, d_{GI}) \in \Delta_{\alpha,\beta}(\mathcal{P}_{\ell,u})$ *when the thresholds* a, b, c, d *are selected as follows:*

$$\begin{aligned} a &= |\log \beta| + \log K, \quad d = |\log \beta| + \log(uK), \\ b &= |\log \alpha| + \log K, \quad c = |\log \alpha| + \log((K-\ell)K). \end{aligned} \tag{2.14}$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_{\ell,u}$ and $a, b, c, d > 0$. Observe that $T_{GI} \leq \tau_2 \leq \tau_2'$, where

$$\tau_2' = \inf\{n \geq 1 : -\lambda^j(n) \geq a,\ \lambda^k(n) \geq b \ \text{ for every } k \in \mathcal{A}, j \notin \mathcal{A}\}. \tag{2.15}$$

Due to assumption (2.2), $\mathsf{P}_{\mathcal{A}}(\tau_2' < \infty) = 1$, which proves that T_{GI} is also almost surely finite under $\mathsf{P}_{\mathcal{A}}$. We now focus on proving the bound in (2.13) for the familywise type-II error probability, since the corresponding result for the familywise type-I error can be shown similarly. From Boole's inequality we have

$$\mathsf{P}_{\mathcal{A}}(d_{GI} \lesssim \mathcal{A}) = \mathsf{P}_{\mathcal{A}}\left(\bigcup_{k\in\mathcal{A}} \{d_{GI}^k = 0\}\right) \leq \sum_{k\in\mathcal{A}} \mathsf{P}_{\mathcal{A}}\left(d_{GI}^k = 0\right). \tag{2.16}$$

Fix $k \in \mathcal{A}$. Whenever the gap-intersection rule mistakenly accepts H_0^k, either the event $\Gamma_k := \{\lambda^k(T_{GI}) \leq -a\}$ occurs (which is the case when stopping at τ_1 or τ_2), or there is at least one $j \notin \mathcal{A}$ such that the event $\Gamma_{k,j} := \{\lambda^j(T_{GI}) - \lambda^k(T_{GI}) \geq d\}$ occurs (which is the case when stopping at τ_3). Therefore,

$$\{d_{GI}^k = 0\} \subset \Gamma_k \cup (\cup_{j\notin\mathcal{A}} \Gamma_{k,j}),$$

and from Boole's inequality we have

$$\mathsf{P}_{\mathcal{A}}(d_{GI}^k = 0) \leq \mathsf{P}_{\mathcal{A}}(\Gamma_k) + \sum_{j\notin\mathcal{A}} \mathsf{P}_{\mathcal{A}}\left(\Gamma_{k,j}\right).$$

Identically to (2.8) we can show that for every $j \notin \mathcal{A}$ we have $\mathsf{P}_{\mathcal{A}}\left(\Gamma_{k,j}\right) \leq e^{-d}$. Moreover, if we set $\mathcal{C} = A\backslash\{k\}$ (note that $C \notin \mathcal{P}_{\ell,u}$, but this does not affect our argument), then $\lambda^{\mathcal{A},\mathcal{C}} = \lambda^k$ and from Wald's likelihood ratio identity we have

$$\mathsf{P}_{\mathcal{A}}(\Gamma_k) = \mathsf{E}_{\mathcal{C}}\left[\exp\{\lambda^{\mathcal{A},\mathcal{C}}(T_{GI})\}; \Gamma_k\right] = \mathsf{E}_{\mathcal{C}}\left[\exp\{\lambda^k(T_{GI})\}; \Gamma_k\right] \leq e^{-a}.$$

Thus,

$$\mathsf{P}_{\mathcal{A}}(d_{GI}^k = 0) \leq e^{-a} + (K - |\mathcal{A}|)e^{-d},$$

which together with (2.16) yields

$$\mathsf{P}_{\mathcal{A}}(d_{GI} \lesssim \mathcal{A}) \leq |\mathcal{A}|(e^{-a} + |\mathcal{A}^c|e^{-d}) \leq \frac{|\mathcal{A}|}{K}(Ke^{-a}) + \frac{|\mathcal{A}^c|}{K}(uKe^{-d}).$$

Therefore, if the thresholds are selected according to (2.14), then $Ke^{-a} = \beta$ and $uKe^{-d} = \beta$, which implies

that

$$\mathsf{P}_{\mathcal{A}}(d_{GI} \lesssim \mathcal{A}) \leq \frac{|\mathcal{A}|}{K}\beta + \frac{|\mathcal{A}^c|}{K}\beta = \beta,$$

and the proof is complete. □

2.4 Computation of familywise error probabilities via importance sampling

The threshold specifications in (2.6) and (2.14) guarantee the desired error control for the gap rule and gap-intersection rule respectively, however they can be very conservative. In practice, it is preferable to use Monte Carlo simulation to determine the thresholds that equate (at least, approximately) the *maximal* familywise type I and type II error probabilities to the corresponding target levels α and β, respectively. Note that this needs to be done offline, before the implementation of the procedure.

When α and β are very small, the corresponding errors are "rare events" and plain Monte Carlo will not be efficient. For this reason, in this section we propose a Monte Carlo approach based on *importance sampling* for the efficient computation of the familywise error probabilities of the proposed multiple testing procedures.

To be more specific, let $\mathcal{A} \subset [K]$ be the true subset of signals and consider the computation of the familywise type I error probability, $\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d)$, of an arbitrary multiple testing procedure, (T, d). The idea of importance sampling is to find a probability measure $\mathsf{P}^*_{\mathcal{A}}$, under which the stopping time T is finite almost surely, and compute the desired probability by estimating (via plain Monte Carlo) the expectation in the right-hand side of the following identity:

$$\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d) = \mathsf{E}^*_{\mathcal{A}}\left[(\Lambda^*_{\mathcal{A}})^{-1}; \mathcal{A} \lesssim d\right],$$

which is obtained by an application of Wald's likelihood ratio identity. Here, we denote by $\Lambda^*_{\mathcal{A}}$ the likelihood ratio of $\mathsf{P}^*_{\mathcal{A}}$ against $\mathsf{P}_{\mathcal{A}}$ at time T, i.e.,

$$\Lambda^*_{\mathcal{A}} = \frac{d\mathsf{P}^*_{\mathcal{A}}}{d\mathsf{P}_{\mathcal{A}}}(\mathcal{F}_T),$$

and by $\mathsf{E}^*_{\mathcal{A}}$ the expectation under $\mathsf{P}^*_{\mathcal{A}}$. The proposal distribution $\mathsf{P}^*_{\mathcal{A}}$ should be selected such that $\Lambda^*_{\mathcal{A}}$ is "large" on the event $\{\mathcal{A} \lesssim d\}$ and "small" on its complement. This intuition will guide us in the selection of $\mathsf{P}^*_{\mathcal{A}}$ for the proposed rules.

For the gap rule (T_G, d_G) we suggest the proposal distribution to be a uniform mixture over the set of distributions $\{\mathsf{P}_{\mathcal{A}\cup\{j\}\setminus\{k\}}, k \in \mathcal{A}, j \notin \mathcal{A}\}$, i.e.,

$$\mathsf{P}^G_{\mathcal{A}} := \frac{1}{|\mathcal{A}|\,|\mathcal{A}^c|} \sum_{k\in\mathcal{A}} \sum_{j\notin\mathcal{A}} \mathsf{P}_{\mathcal{A}\cup\{j\}\setminus\{k\}}, \tag{2.17}$$

whose likelihood ratio against $\mathsf{P}_{\mathcal{A}}$ at time T_G is

$$\Lambda^G_{\mathcal{A}} := \frac{1}{|\mathcal{A}|\,|\mathcal{A}^c|} \sum_{k\in\mathcal{A}} \sum_{j\notin\mathcal{A}} \exp\{\lambda^j(T_G) - \lambda^k(T_G)\}.$$

Then, on the event $\{\mathcal{A} \lesssim d_G\}$ there exists some $k \in \mathcal{A}$ and $j \notin \mathcal{A}$ such that $\lambda^j(T_G) - \lambda^k(T_G) \geq c$, which leads to a large value for $\Lambda^G_{\mathcal{A}}$. On the other hand, on the complement of $\{\mathcal{A} \lesssim d_G\}$, $\{d_G = \mathcal{A}\}$, we have $\lambda^j(T_G) - \lambda^k(T_G) \leq -c$ for every $k \in \mathcal{A}, j \notin \mathcal{A}$, which leads to a value of $\Lambda^G_{\mathcal{A}}$ close to 0.

For the intersection rule (T_I, d_I) we suggest the proposal distribution to be a uniform mixture over $\{\mathsf{P}_{\mathcal{A}\cup\{j\}}, j \notin \mathcal{A}\}$, i.e.,

$$\mathsf{P}^I_{\mathcal{A}} := \frac{1}{|\mathcal{A}^c|} \sum_{j\notin\mathcal{A}} \mathsf{P}_{\mathcal{A}\cup\{j\}}, \tag{2.18}$$

whose likelihood ratio against $\mathsf{P}_{\mathcal{A}}$ at time T_I takes the form

$$\Lambda^I_{\mathcal{A}} := \frac{1}{|\mathcal{A}^c|} \sum_{j\notin\mathcal{A}} \exp\{\lambda^j(T_I)\}.$$

Note that on the event $\{\mathcal{A} \lesssim d_I\}$ there exists some $j \notin A$ such that $\lambda^j(T_I) \geq b$, which results in a large value for $\Lambda^I_{\mathcal{A}}$. On the other hand, on the complement of $\{\mathcal{A} \lesssim d_I\}$ we have $\lambda^j(T_I) \leq -a$ for every $j \notin \mathcal{A}$, which results in a value of $\Lambda^I_{\mathcal{A}}$ close to 0.

Finally, for the gap-intersection rule we suggest to use $\mathsf{P}^I_{\mathcal{A}}$, the same proposal distribution as in the intersection rule, when $\ell < |\mathcal{A}| < u$. In the boundary case, i.e. $|\mathcal{A}| = \ell$ or $|\mathcal{A}| = u$, we propose the following mixture of $\mathsf{P}^G_{\mathcal{A}}$ and $\mathsf{P}^I_{\mathcal{A}}$:

$$\mathsf{P}^{GI}_{\mathcal{A}} := \frac{|\mathcal{A}|}{1+|\mathcal{A}|}\,\mathsf{P}^G_{\mathcal{A}} + \frac{1}{1+|\mathcal{A}|}\,\mathsf{P}^I_{\mathcal{A}}.$$

In Section 2.6 we apply the proposed simulation approach for the specification of non-conservative thresholds in the case of identical, symmetric hypotheses with Gaussian i.i.d. data. We also refer to [65] for an analysis of these importance sampling estimators.

2.5 Asymptotic optimality in the i.i.d. setup

From now on, we assume that, for each stream $k \in [K]$, the observations $\{X_n^k, n \in \mathbb{N}\}$ are independent random variables with common density f_i^k with respect to a σ-finite measure μ^k under P_i^k, $i = 0, 1$, such that the Kullback—-Leibler information numbers

$$D_0^k := \int \log\left(\frac{f_0^k}{f_1^k}\right) f_0^k d\mu^k, \quad D_1^k := \int \log\left(\frac{f_1^k}{f_0^k}\right) f_1^k d\mu^k$$

are both positive and finite. As a result, for each $k \in [K]$ the log-likelihood ratio process in the k^{th} stream, defined in (2.1), takes the form

$$\lambda^k(n) = \sum_{j=1}^{n} \log \frac{f_1^k(X_j^k)}{f_0^k(X_j^k)}, \quad n \in \mathbb{N},$$

and it is a random walk with drift D_1^k under P_1^k and $-D_0^k$ under P_0^k.

Our goal in this section is to show that the proposed multiple testing procedures in Section 2.3 are asymptotically optimal. Our strategy for proving this is first to establish a *non-asymptotic* lower bound on the minimum possible expected sample size in $\Delta_{\alpha,\beta}(\mathcal{P})$ for some arbitrary class $\mathcal{P}$, and then show that this lower bound is attained by the gap rule when $\mathcal{P} = \mathcal{P}_m$ and by the gap-intersection rule when $\mathcal{P} = \mathcal{P}_{\ell,u}$ as $\alpha, \beta \to 0$.

2.5.1 A lower bound on the optimal performance

In order to state the lower bound on the optimal performance, we introduce the function

$$\varphi(x, y) := x \log\left(\frac{x}{1-y}\right) + (1-x) \log\left(\frac{1-x}{y}\right), \quad x, y \in (0, 1), \tag{2.19}$$

and for any subsets $\mathcal{C}, \mathcal{A} \subset [K]$ such that $\mathcal{C} \neq \mathcal{A}$ we set

$$\gamma_{\mathcal{A},\mathcal{C}}(\alpha, \beta) := \begin{cases} \varphi(\alpha, \beta), & \text{if } \mathcal{C} \setminus \mathcal{A} \neq \emptyset,\ \mathcal{A} \setminus \mathcal{C} = \emptyset, \\ \varphi(\beta, \alpha), & \text{if } \mathcal{C} \setminus \mathcal{A} = \emptyset,\ \mathcal{A} \setminus \mathcal{C} \neq \emptyset, \\ \varphi(\alpha, \beta) \vee \varphi(\beta, \alpha), & \text{otherwise.} \end{cases}$$

Theorem 2.3. *For any class $\mathcal{P}$, $\mathcal{A} \in \mathcal{P}$ and $\alpha, \beta \in (0, 1)$ such that $\alpha + \beta < 1$ we have*

$$\inf_{(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P})} \mathsf{E}_{\mathcal{A}}[T] \geq \max_{\mathcal{C} \in \mathcal{P}, \mathcal{C} \neq \mathcal{A}} \frac{\gamma_{\mathcal{A},\mathcal{C}}(\alpha, \beta)}{\sum_{k \in \mathcal{A} \setminus \mathcal{C}} D_1^k + \sum_{k \in \mathcal{C} \setminus \mathcal{A}} D_0^k}. \tag{2.20}$$

Proof. Fix $(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P})$ and $\mathcal{A} \in \mathcal{P}$. Without loss of generality, we assume that $\mathsf{E}_{\mathcal{A}}[T] < \infty$. For any $\mathcal{C} \in \mathcal{P}$ such that $\mathcal{C} \neq \mathcal{A}$, the log-likelihood ratio process $\lambda^{\mathcal{A},\mathcal{C}}$, defined in (2.3), is a random walk under $\mathsf{P}_{\mathcal{A}}$ with drift equal to

$$\mathsf{E}_{\mathcal{A}}[\lambda^{\mathcal{A},\mathcal{C}}(1)] = \sum_{k \in \mathcal{A}\setminus\mathcal{C}} D_1^k + \sum_{k \in \mathcal{C}\setminus\mathcal{A}} D_0^k,$$

since each λ^k is a random walk with drift D_1^k under P_1^k and $-D_0^k$ under P_0^k. Thus, from Wald's identity it follows that

$$\mathsf{E}_{\mathcal{A}}[T] = \frac{\mathsf{E}_{\mathcal{A}}[\lambda^{\mathcal{A},\mathcal{C}}(T)]}{\sum_{k \in \mathcal{A}\setminus\mathcal{C}} D_1^k + \sum_{k \in \mathcal{C}\setminus\mathcal{A}} D_0^k},$$

and it suffices to show that for any $\mathcal{C} \in \mathcal{P}$ such that $\mathcal{C} \neq \mathcal{A}$ we have

$$\mathsf{E}_{\mathcal{A}}[\lambda^{\mathcal{A},\mathcal{C}}(T)] \geq \gamma_{\mathcal{A},\mathcal{C}}(\alpha,\beta). \tag{2.21}$$

Suppose that $\mathcal{C} \setminus \mathcal{A} \neq \emptyset$ and let $j \in \mathcal{C} \setminus \mathcal{A}$. Then, from Lemma 2.3 in the Section 2.8 we have

$$\mathsf{E}_{\mathcal{A}}\left[\lambda^{\mathcal{A},\mathcal{C}}(T)\right] = \mathsf{E}_{\mathcal{A}}\left[\log \frac{d\mathsf{P}_{\mathcal{A}}}{d\mathsf{P}_{\mathcal{C}}}(\mathcal{F}_T)\right] \geq \varphi\left(\mathsf{P}_{\mathcal{A}}(d^j = 1), \mathsf{P}_{\mathcal{C}}(d^j = 0)\right).$$

By the definition of $\Delta_{\alpha,\beta}(\mathcal{P})$, we have $\mathsf{P}_{\mathcal{A}}(d^j = 1) \leq \alpha$ and $\mathsf{P}_{\mathcal{C}}(d^j = 0) \leq \beta$. Since the function $\varphi(x,y)$ is decreasing on the set $\{(x,y) : x + y \leq 1\}$, and by assumption $\alpha + \beta \leq 1$, we conclude that if $\mathcal{C} \setminus \mathcal{A} \neq \emptyset$, then

$$\mathsf{E}_{\mathcal{A}}[\lambda^{\mathcal{A},\mathcal{C}}(T)] \geq \varphi(\alpha,\beta).$$

With a symmetric argument we can show that if $\mathcal{A} \setminus \mathcal{C} \neq \emptyset$, then

$$\mathsf{E}_{\mathcal{A}}[\lambda^{\mathcal{A},\mathcal{C}}(T)] \geq \varphi(\beta,\alpha).$$

The two last inequalities imply (2.21), and this completes the proof. □

Remark 2.1. *By the definition of φ in* (2.19)*, we have*

$$\varphi(\alpha,\beta) = |\log\beta|\,(1 + o(1)), \quad \varphi(\beta,\alpha) = |\log\alpha\,|(1 + o(1)) \tag{2.22}$$

as $\alpha, \beta \to 0$ at arbitrary rates.

2.5.2 Asymptotic optimality of the proposed schemes

In what follows, we assume that for each stream $k \in [K]$ we have:

$$\int \left(\log \left(\frac{f_0^k}{f_1^k} \right) \right)^2 f_i^k d\mu^k < \infty, \quad i = 0, 1. \tag{2.23}$$

Although this assumption is not necessary for the asymptotic optimality of the proposed rules to hold, it will allow us to use Lemma 2.4 in the Section 2.8 and obtain valuable insights regarding the effect of prior information on the optimal performance. Moreover, for each subset $\mathcal{A} \subset [K]$ we set:

$$\eta_1^{\mathcal{A}} := \min_{k \in \mathcal{A}} D_1^k, \qquad \eta_0^{\mathcal{A}} := \min_{j \notin \mathcal{A}} D_0^j,$$

and, following the convention that the minimum over the empty set is ∞, we define: $\eta_1^{\emptyset} = \eta_0^{[K]} := \infty$.

Known number of signals

We will first show that the gap rule, defined in (2.4), is asymptotically optimal with respect to class $\mathcal{P}_m$, where $1 \leq m \leq K-1$. In order to do so, we start with an upper bound on the expected sample size of this procedure.

Lemma 2.1. *Suppose that assumption* (2.23) *holds. Then, for any $\mathcal{A} \in \mathcal{P}_m$, as $c \to \infty$ we have*

$$\mathsf{E}_{\mathcal{A}}[T_G] \leq \frac{c}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}} + O\left(m(K-m)\sqrt{c}\right).$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_m$. For any $c > 0$ we have $T_G \leq T_G'$, where T_G' is defined in (2.7), and it is the first time that all $m(K-m)$ processes of the form $\lambda^k - \lambda^j$ with $k \in \mathcal{A}$ and $j \notin \mathcal{A}$ exceed c. Due to condition (2.23), each $\lambda^k - \lambda^j$ with $k \in \mathcal{A}$ and $j \notin \mathcal{A}$ is a random walk under $\mathsf{P}_{\mathcal{A}}$ with positive drift $D_1^k + D_0^j$ and finite second moment. Therefore, from Lemma 2.4 it follows that as $c \to \infty$:

$$\mathsf{E}_{\mathcal{A}}[T_G'] \leq c \left(\min_{k \in \mathcal{A}, j \notin \mathcal{A}} (D_1^k + D_0^j) \right)^{-1} + O\left(m(K-m)\sqrt{c}\right),$$

and this completes the proof, since $\min_{k \in \mathcal{A}, j \notin \mathcal{A}} (D_1^k + D_0^j) = \eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}$. □

The next theorem establishes the asymptotic optimality of the gap rule.

Theorem 2.4. *Suppose assumption* (2.23) *holds and let the threshold c in the gap rule be selected according*

to (2.6). *Then for every* $\mathcal{A} \in \mathcal{P}_m$, *we have as* $\alpha, \beta \to 0$

$$\mathsf{E}_{\mathcal{A}}[T_G] \sim \frac{|\log(\alpha \wedge \beta)|}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}} \sim \inf_{(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P}_m)} \mathsf{E}_{\mathcal{A}}[T].$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_m$. If thresholds are selected according to (2.6), then from Lemma 2.1 it follows that as $\alpha, \beta \to 0$

$$\mathsf{E}_{\mathcal{A}}[T_G] \leq \frac{|\log(\alpha \wedge \beta)|}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}} + O\left(m(K-m)\sqrt{|\log(\alpha \wedge \beta)|}\right). \tag{2.24}$$

Therefore, it suffices to show that the lower bound in Theorem 2.3 agrees with the upper bound in (2.24) in the first-order term as $\alpha, \beta \to 0$. To see this, note that for any $\mathcal{C} \in \mathcal{P}_m$ such that $\mathcal{C} \neq \mathcal{A}$ we have $\mathcal{C} \setminus \mathcal{A} \neq \emptyset$ *and* $\mathcal{A} \setminus \mathcal{C} \neq \emptyset$, and consequently

$$\gamma_{\mathcal{A},\mathcal{C}}(\alpha, \beta) = \varphi(\alpha, \beta) \vee \varphi(\beta, \alpha).$$

This means that the numerator in (2.20) does not depend on $\mathcal{C}$. Moreover, if we restrict our attention to subsets in $\mathcal{P}_m$ that differ from $\mathcal{A}$ in two streams, i.e., subsets of the form $\mathcal{C} = \mathcal{A} \cup \{j\} \setminus \{k\}$ for some $k \in \mathcal{A}$ and $j \notin \mathcal{A}$, for which

$$\sum_{i \in \mathcal{A} \setminus \mathcal{C}} D_1^i + \sum_{i \in \mathcal{C} \setminus \mathcal{A}} D_0^i = D_1^k + D_0^j,$$

then we have

$$\min_{\mathcal{C} \in \mathcal{P}_m, \mathcal{C} \neq \mathcal{A}} \left[\sum_{i \in \mathcal{A} \setminus \mathcal{C}} D_1^i + \sum_{i \in \mathcal{C} \setminus \mathcal{A}} D_0^i \right] \leq \min_{k \in \mathcal{A}, j \notin \mathcal{A}} \left[D_1^k + D_0^j \right] = \eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}.$$

By the last inequality and Theorem 2.3 we obtain the following non-asymptotic lower bound, which holds for any α, β such that $\alpha + \beta < 1$:

$$\inf_{(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P}_m)} \mathsf{E}_{\mathcal{A}}[T] \geq \frac{\max\{\varphi(\alpha, \beta), \varphi(\beta, \alpha)\}}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}}.$$

By (2.22), we have as $\alpha, \beta \to 0$

$$\max\{\varphi(\alpha, \beta), \varphi(\beta, \alpha)\} = |\log(\alpha \wedge \beta)|\,(1 + o(1)).$$

Consequently,

$$\inf_{(T,d) \in \Delta_{\alpha,\beta}(\mathcal{P}_m)} \mathsf{E}_{\mathcal{A}}(T) \geq \frac{|\log(\alpha \wedge \beta)|}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}}\,(1 + o(1)),$$

which completes the proof. □

Remark 2.2. *It is interesting to consider the special case of identical hypotheses, in which* $f_1^k = f_1$ *and* $f_0^k = f_0$*, and consequently* $D_1^k = D_1$ *and* $D_0^k = D_0$ *for every* $k \in [K]$*. Then,* $\eta_1^{\mathcal{A}} = D_1$ *and* $\eta_0^{\mathcal{A}} = D_0$ *for every* $\mathcal{A} \subset [K]$*, and from Theorem 2.4 it follows that the first-order asymptotic approximation to the expected sample size of the gap rule (as well as to the optimal expected sample size within* $\Delta_{\alpha,\beta}(\mathcal{P}_m)$*),* $|\log(\alpha \wedge \beta)|/(D_1 + D_0)$*, is independent of the number of signals,* m*. We should stress that this does not mean that the actual performance of the gap rule is independent of* m*. Indeed, the second term in the right-hand side of* (2.24) *suggests that the smaller* $m(K-m)$ *is, i.e., the further away the proportion of signals* m/K *is from* $1/2$*, the smaller the expected sample size of the gap rule will be. This intuition will be corroborated by the simulation study in Section 2.6 (see Figure 2.2).*

Lower and upper bounds on the number of signals

We will now show that the gap-intersection rule, defined in (2.12), is asymptotically optimal with respect to class $\mathcal{P}_{\ell,u}$ for some $0 \le \ell < u \le K$. As before, we start with establishing an upper bound on the expected sample size of this rule.

Lemma 2.2. *Suppose that assumption* (2.23) *holds. Then, for any* $\mathcal{A} \in \mathcal{P}_{\ell,u}$ *we have as* $a, b, c, d \to \infty$

$$\mathsf{E}_{\mathcal{A}}[T_{GI}] \le \begin{cases} \max\left\{a/\eta_0^{\mathcal{A}}\ ,\ c/(\eta_0^{\mathcal{A}} + \eta_1^{\mathcal{A}})\right\}(1 + o(1)) & \text{if } |\mathcal{A}| = \ell \\ \max\left\{a/\eta_0^{\mathcal{A}}\ ,\ b/\eta_1^{\mathcal{A}}\right\} + O(K\sqrt{a \vee b}) & \text{if } \ell < |\mathcal{A}| < u \\ \max\left\{b/\eta_1^{\mathcal{A}}\ ,\ d/(\eta_0^{\mathcal{A}} + \eta_1^{\mathcal{A}})\right\}(1 + o(1)) & \text{if } |\mathcal{A}| = u \end{cases}$$

Furthermore, if $c - a = O(1)$ *and* $d - b = O(1)$*, then*

$$\mathsf{E}_{\mathcal{A}}[T_{GI}] \le \begin{cases} a/\eta_0^{\mathcal{A}} + O((K-\ell)\sqrt{a}) & \text{if } |\mathcal{A}| = \ell \\ b/\eta_1^{\mathcal{A}} + O(u\sqrt{b}) & \text{if } |\mathcal{A}| = u \end{cases} \tag{2.25}$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_{\ell,u}$. By the definition of the stopping time T_{GI},

$$\mathsf{E}_{\mathcal{A}}[T_{GI}] \le \min\left\{\mathsf{E}_{\mathcal{A}}[\tau_1], \mathsf{E}_{\mathcal{A}}[\tau_2], \mathsf{E}_{\mathcal{A}}[\tau_3]\right\}.$$

Suppose first $\ell < |\mathcal{A}| < u$ and observe that $\tau_2 \le \tau_2'$, where τ_2' is defined in (2.15). Under condition (2.23), for every $k \in \mathcal{A}$ and $j \notin \mathcal{A}$, $-\lambda^j$ and λ^k are random walks with finite second moments and positive drifts

D_0^j and D_1^k, respectively. Therefore, from Lemma 2.4 we have that

$$\mathsf{E}_{\mathcal{A}}[\tau_2'] \leq \max\left\{a/\eta_0^{\mathcal{A}}\ ,\ b/\eta_1^{\mathcal{A}}\right\} + O(K\sqrt{a \vee b}).$$

Suppose now that $|\mathcal{A}| = \ell$ and observe that $\tau_1 \leq \tau_1'$, where

$$\tau_1' := \inf\{n \geq 1 : -\lambda^j(n) \geq a,\ \lambda^k(n) - \lambda^j(n) \geq c \ \text{ for every } k \in \mathcal{A}, j \notin \mathcal{A}\},$$

where $-\lambda^j$ and $\lambda^k - \lambda^j$ are random walks with finite second moments and positive drifts D_0^j and $D_1^k + D_0^j$, respectively. The result follows again from an application of Lemma 2.4. If in addition we have that $c - a = O(1)$, then $\tau_1 \leq \tau_1''$, where

$$\tau_1'' := \inf\{n \geq 1 : -\lambda^j(n) \geq a,\ \lambda^k(n) \geq c - a \ \text{ for every } k \in \mathcal{A}, j \notin \mathcal{A}\}.$$

Therefore, the second part of the lemma follows again from an application of Lemma 2.4. □

The next theorem establishes the asymptotic optimality of the gap-intersection rule.

Theorem 2.5. *Suppose that assumption* (2.23) *holds and let the thresholds in the gap-intersection rule be selected according to* (2.14)*. Then for any $\mathcal{A} \in \mathcal{P}_{\ell,u}$, we have as $\alpha, \beta \to 0$*

$$\begin{aligned}
\mathsf{E}_{\mathcal{A}}[T_{GI}] \ &\sim \inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{\ell,u})} \mathsf{E}_{\mathcal{A}}[T] \\
&\sim \begin{cases}
\max\left\{|\log\beta|/\eta_0^{\mathcal{A}}\ ,\ |\log\alpha|/(\eta_0^{\mathcal{A}} + \eta_1^{\mathcal{A}})\right\} & \text{if } \ |\mathcal{A}| = \ell \\
\max\left\{|\log\beta|/\eta_0^{\mathcal{A}}\ ,\ |\log\alpha|/\eta_1^{\mathcal{A}}\right\} & \text{if } \ \ell < |\mathcal{A}| < u \\
\max\left\{|\log\alpha|/\eta_1^{\mathcal{A}}\ ,\ |\log\beta|/(\eta_0^{\mathcal{A}} + \eta_1^{\mathcal{A}})\right\} & \text{if } \ |\mathcal{A}| = u
\end{cases}.
\end{aligned}$$

Proof. Fix $\mathcal{A} \in \mathcal{P}_{\ell,u}$. We will prove the result only in the case that $|\mathcal{A}| = \ell$, as the other two cases can be proved similarly. If thresholds are selected according to (2.14), then from Lemma 2.2 it follows that

$$\mathsf{E}_{\mathcal{A}}[T_{GI}] \leq \max\left\{\frac{|\log\beta|}{\eta_0^{\mathcal{A}}}\ ,\ \frac{|\log\alpha|}{\eta_0^{\mathcal{A}} + \eta_1^{\mathcal{A}}}\right\}(1 + o(1)).$$

Thus, it suffices to show that this asymptotic upper bound agrees asymptotically, *up to a first order*, with the lower bound in Theorem 2.3. Indeed, if $\mathcal{C}$ is a subset in $\mathcal{P}_{\ell,u}$ that has one more stream than $\mathcal{A}$, i.e.,

$\mathcal{C} = \mathcal{A} \cup \{j\}$ for some $j \notin \mathcal{A}$, then

$$\frac{\gamma_{\mathcal{A},\mathcal{C}}(\alpha,\beta)}{\sum_{i\in\mathcal{A}\setminus\mathcal{C}} D_1^i + \sum_{i\in\mathcal{C}\setminus\mathcal{A}} D_0^i} = \frac{\varphi(\alpha,\beta)}{D_0^j}.$$

Further, consider $\mathcal{C} = \mathcal{A} \cup \{j\}/\{k\} \in \mathcal{P}_{\ell,u}$ for some $k \in \mathcal{A}$ and $j \notin \mathcal{A}$, then

$$\frac{\gamma_{\mathcal{A},\mathcal{C}}(\alpha,\beta)}{\sum_{i\in\mathcal{A}\setminus\mathcal{C}} D_1^i + \sum_{i\in\mathcal{C}\setminus\mathcal{A}} D_0^i} = \frac{\max\{\varphi(\alpha,\beta),\varphi(\beta,\alpha)\}}{D_1^k + D_0^j}.$$

Therefore, from (2.3) it follows that for every α, β such that $\alpha + \beta < 1$

$$\begin{aligned}\inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{\ell,u})} \mathsf{E}_{\mathcal{A}}[T] &\geq \max_{k\in\mathcal{A}, j\notin\mathcal{A}} \max\left\{ \frac{\varphi(\alpha,\beta)}{D_0^j} \ , \ \frac{\max\{\varphi(\alpha,\beta),\varphi(\beta,\alpha)\}}{D_1^k + D_0^j} \right\} \\ &= \max\left\{ \frac{\varphi(\alpha,\beta)}{\eta_0^{\mathcal{A}}}, \ \frac{\varphi(\beta,\alpha)}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}} \right\}.\end{aligned}$$

From (2.22) it follows that as $\alpha, \beta \to 0$

$$\inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{l,u})} \mathsf{E}_{\mathcal{A}}[T] \geq \max\left\{ \frac{|\log\beta|}{\eta_0^{\mathcal{A}}}, \ \frac{|\log\alpha|}{\eta_1^{\mathcal{A}} + \eta_0^{\mathcal{A}}} \right\} (1 + o(1)),$$

which completes the proof. □

2.5.3 The case of no prior information

Recall that when we set $\ell = 0$ and $u = K$, the gap-intersection rule reduces to the intersection rule, defined in (2.9). Therefore, setting $\ell = 0$ and $u = K$ in Theorem 2.5 we immediately obtain that the intersection rule is asymptotically optimal in the case of no prior information, i.e., with respect to class $\mathcal{P}_{0,K}$; this is itself a new result to the best of our knowledge. However, a more surprising corollary of Theorem 2.5 is that the intersection rule, which does not use any prior information, is asymptotically optimal even if bounds on the number of signals are available, when the following conditions are satisfied:

(i) the error probabilities are of the same order of magnitude, in the sense that $|\log\alpha| \sim |\log\beta|$,

(ii) the hypotheses are identical and symmetric, in the sense that $D_1^k = D_0^k = D$ for every $k \in [K]$.

On the other hand, a comparison with Theorem 2.4 reveals that, even in this special case, the intersection rule is never asymptotically optimal when the exact umber of signals is known in advance, in which case it requires roughly *twice* as many observations on average as the gap rule for the same precision level. The following corollary summarizes these observations.

Corollary 2.1. *Suppose that assumption* (2.23) *holds and that the thresholds in the intersection rule are selected according to* (2.10)*. Then, for any* $\mathcal{A} \subset [K]$ *we have as* $\alpha, \beta \to 0$

$$\mathsf{E}_{\mathcal{A}}[T_I] \leq \max\left\{\frac{|\log\alpha|}{\eta_1^{\mathcal{A}}}, \frac{|\log\beta|}{\eta_0^{\mathcal{A}}}\right\} + O(K\sqrt{|\log(\alpha\wedge\beta)|}). \tag{2.26}$$

Further, the intersection rule is asymptotically optimal in the class $\Delta_{\alpha,\beta}(\mathcal{P}_{0,K})$*, i.e., as* $\alpha, \beta \to 0$

$$\mathsf{E}_{\mathcal{A}}[T_I] \;\sim\; \max\left\{\frac{|\log\alpha|}{\eta_1^{\mathcal{A}}}, \frac{|\log\beta|}{\eta_0^{\mathcal{A}}}\right\} \;\sim\; \inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{0,K})} \mathsf{E}_{\mathcal{A}}[T].$$

In the special case that $|\log\alpha| \sim |\log\beta|$ *and* $D_1^k = D_0^k = D$ *for every* $k \in [K]$*,*

$$\begin{aligned}
\mathsf{E}_{\mathcal{A}}[T_I] &\sim \frac{|\log\alpha|}{D} \;\sim\; \inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{\ell,u})} \mathsf{E}_{\mathcal{A}}[T] \quad \text{for every } \mathcal{A} \in \mathcal{P}_{\ell,u},\\
\mathsf{E}_{\mathcal{A}}[T_I] &\sim \frac{|\log\alpha|}{D} \;\sim\; 2\inf_{(T,d)\in\Delta_{\alpha,\beta}(\mathcal{P}_{m})} \mathsf{E}_{\mathcal{A}}[T] \quad \text{for every } \mathcal{A} \in \mathcal{P}_{m},
\end{aligned}$$

for every $0 \leq \ell < u \leq K$ *and* $1 \leq m \leq K-1$*.*

Remark 2.3. *Corollary 2.1 implies that, in the special symmetric case that* $|\log\alpha| \sim |\log\beta|$ *and* $D_1^k = D_0^k = D$*, prior lower and upper bounds on the true number of signals do not improve the optimal expected sample size up to a first-order asymptotic approximation. However, a comparison between the second-order terms in* (2.25) *and* (2.26) *suggests that such prior information does improve the optimal performance, an intuition that will be corroborated by the simulation study in Section 2.6 (see Figure 2.2).*

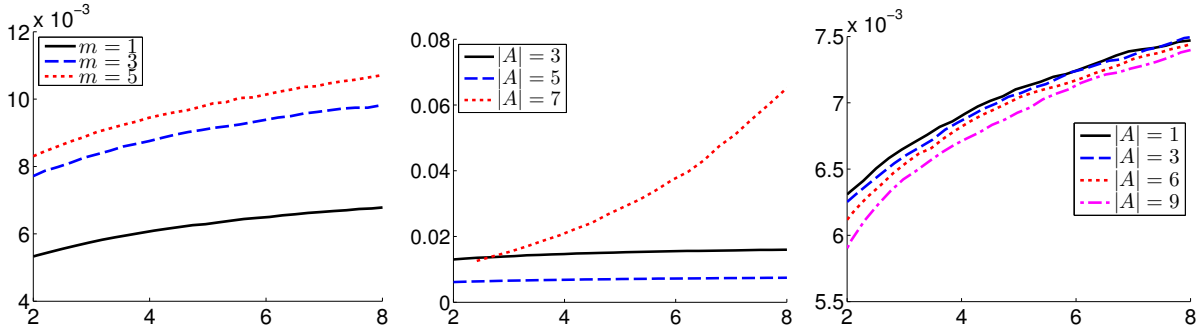
Remark 2.4. *In addition to the intersection rule, De and Baron [17] proposed the "incomplete rule",* $(T_{\max}, d_{\max})$*, which is defined as*

$$T_{\max} := \max\{\sigma_1, \ldots, \sigma_K\} \;\; \text{and} \;\; d_{\max} := (d_{\max}^1, \ldots, d_{\max}^K),$$

where for every $k \in [K]$ *we have*

$$\sigma_k := \inf\left\{n \geq 1 : \lambda^k(n) \notin (-a, b)\right\}, \quad d_{\max}^k := \begin{cases} 1, & \text{if } \lambda^k(\sigma_k) \geq b \\ 0, & \text{if } \lambda^k(\sigma_k) \leq -a \end{cases}. \tag{2.27}$$

According to this rule, each stream is sampled until the corresponding test statistic exits the interval $(-a, b)$*, independently of the other streams. It is clear that, for the same thresholds* a *and* b*,* $T_{\max} \leq T_I$*. Moreover, with a direct application of Boole's inequality, as in De and Baron [17], it follows that selecting the thresholds*



(a) Gap rule (b) Gap-intersection rule (c) Intersection rule

Figure 2.1: The x-axis is $|\log_{10}(\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d))|$. The y-axis is the relative error of the estimate of the familywise type-I error, $\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d)$, that is the ratio of the standard deviation of the estimate over the estimate itself. Each curve is computed based on $100,000$ realizations.

according to (2.10) *guarantees the desired error control for the incomplete rule. Therefore, Corollary 2.1 remains valid if we replace the intersection rule with the incomplete rule.*

2.6 Simulation study

2.6.1 Description

In this section we present a simulation study whose goal is to corroborate the asymptotic results and insights of Section 2.5 in the symmetric case described in Corollary 2.1. Thus, we set $K = 10$ and let $f_i^k = \mathcal{N}(\theta_i, 1)$ for each $k \in [K]$, $i = 0, 1$, where $\theta_0 = 0, \theta_1 = 0.5$, in which case $D_0^k = D_1^k = D = (1/2)(\theta_1)^2 = 1/8$, and the distribution of λ^k under H_1^k is the same as $-\lambda^k$ under H_0^k. Furthermore, we set $\alpha = \beta$. This is a convenient setup for simulation purposes, since the expected sample size and the two familywise errors of each proposed procedure are the same for all scenarios with the same number of signals, i.e. for all $\mathcal{A}$'s with the same size.

For any user specified level α, we have two ways to determine the critical value of each procedure. First, we can use upper bound on the error probability to compute conservative threshold ((2.6) for the gap rule, and (2.14) for the gap-intersection rule). Second, we can apply the importance sampling technique of Section 2.4 to determine non-conservative threshold, such that the *maximal* familywise type I error probability is controlled *exactly* at level α. As we see in Figure 2.1, the relative errors of the proposed Monte Carlo estimators, even for error probabilities of the order 10^{-8}, are smaller than 1.5% for the gap rule, 8% for the gap-intersection rule, 1% for the intersection rule.

Gap rule

First, we consider the case in which the number of signals is known to be equal to m ($\mathcal{P} = \mathcal{P}_m$) for $m \in \{1, \ldots, 9\}$, and we can apply the corresponding gap rule, defined in (2.4). Due to the symmetry of our setup, the expected sample size $\mathsf{E}_{\mathcal{A}}[T_G]$ and the error probability $\mathsf{P}_{\mathcal{A}}(d_G \neq \mathcal{A})$ are the same for $\mathcal{A} \in \mathcal{P}_m$ and $\mathcal{A} \in \mathcal{P}_{K-m}$; thus, it suffices to consider m in $\{1, \ldots, 5\}$, and an *arbitrary* $\mathcal{A} \in \mathcal{P}_m$ for fixed m.

We start with non-conservative critical value determined by Monte Carlo method. For each $m \in \{1, 3, 5\}$ and some $\mathcal{A} \in \mathcal{P}_m$, we consider α's ranging from 10^{-2} to 10^{-8}. For each such α, we compute the threshold c in the gap-rule that guarantees $\alpha = \max_{\mathcal{A} \in \mathcal{P}_m} \mathsf{P}_{\mathcal{A}}(d_G \neq \mathcal{A})$, and then the expected sample size $\mathsf{E}_{\mathcal{A}}[T_G]$ that corresponds to this threshold. In Figure 2.2a we plot $\mathsf{E}_{\mathcal{A}}[T_G]$ against $|\log_{10}(\alpha)|$ when $m = 1, 3, 5$. In Table 2.1a we present the actual numerical results for $c = 10$.

In Figure 2.2a we also plot the first-order asymptotic approximation to the optimal expected sample size obtained in Theorem 2.4, which in this particular symmetric case takes the form $|\log \alpha|/(2D) = 4|\log \alpha|$. From our asymptotic theory we know that the ratio of $\mathsf{E}_{\mathcal{A}}[T_G]$ over this quantity goes to 1 as $\alpha \to 0$, and this convergence is illustrated in Figure 2.2b.

Further, in Figure 2.3a we present for the case $\mathcal{P} = \mathcal{P}_3$ the expected sample size of the gap rule when its threshold is given by the explicit expression in (2.6), and compare it with the corresponding expected sample size that is obtained with the sharp threshold, which is computed via simulation.

Gap-intersection rule

Second, we consider the case in which the number of signals is known to be between 3 and 7 ($\mathcal{P} = \mathcal{P}_{\ell,u} = \mathcal{P}_{3,7}$), and we can apply the gap-intersection rule, defined in (2.12). Due to the symmetry of the setup and Lemma 2.2, we set $a = b$ and $c = d = b + \log(u) = b + \log(7)$.

As before, we consider α's ranging from 10^{-2} to 10^{-8}. For each such α, we obtain the threshold b such that $\max_{\mathcal{A}} \mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d_{GI}) = \alpha$, where the maximum is taken over $\mathcal{A} \in \mathcal{P}_{\ell,u}$, and then compute the corresponding expected sample size $\mathsf{E}_{\mathcal{A}}[T_{GI}]$ for every $\mathcal{A} \in \mathcal{P}_{\ell,u}$. In Figure 2.2c we plot $\mathsf{E}_{\mathcal{A}}[T_{GI}]$ against $|\log_{10}(\alpha)|$ for $|\mathcal{A}| = 3$ and 5, since by symmetry $\mathsf{E}_{\mathcal{A}}[T_{GI}]$ is the same for $|\mathcal{A}| = k$ and $10 - k$, and the results for $|\mathcal{A}| = 4$ and 5 were too close. This is also evident from Table 2.1b, where we present the numerical results for $b = 10$. In the same graph we also plot the first-order asymptotic approximation to the optimal performance obtained in Theorem 2.5, which in this case is $|\log \alpha|/D = 8|\log \alpha|$. By Theorem 2.5, we know that the ratio of $\mathsf{E}_{\mathcal{A}}[T_{GI}]$ over $8|\log \alpha|$ goes to 1 as $\alpha \to 0$, which is corroborated in Figure 2.2d.

Intersection versus incomplete rule

Finally, we consider the case of no prior information ($\mathcal{P} = \mathcal{P}_{0,10}$), in which we compare the intersection rule with the incomplete rule. This is a special case of the previous setup with $\ell = 0$ and $u = K$, but now the expected sample size (for both schemes) is the same for every subset of signals $\mathcal{A}$, which allows us to plot only one curve for each scheme in Figure 2.2e (*non-conservative* critical value is used). In the same graph we also plot the first-order approximation to the optimal performance, $|\log \alpha|/D = 8|\log \alpha|$, whereas in Figure 2.2f. we plot the corresponding normalized version.

Further, in Figure 2.3b we present the expected sample size of the intersection rule when its threshold is given by the explicit expression in (2.14), and compare it with the corresponding expected sample size that is obtained with the sharp threshold, which is computed via simulation.

2.6.2 Results

There are a number of conclusions that can be drawn from the presented graphs. First of all, from Figure 2.2a it follows that the gap rule performs the best when there are exactly $m = 1$ or 9 signals, whereas its performance is quite similar for $m = 3, 4, 5$. As we mentioned before, this can be explained by the fact that the second term in the right-hand side in (2.24) grows with $m(K - m)$.

Second, from Figure 2.2c we can see that the gap-intersection rule performs better in the boundary cases that there are exactly 3 or 7 signals than in the case of 5 signals, which can be explained by the second order term in (2.25).

Third, from Figure 2.2e we can see that the intersection rule is always better than the incomplete rule, although they share the same prior information.

Fourth, from the graphs in the second column of Figure 2.2 we can see that all curves approach 1, as expected from our asymptotic results; however, the convergence is relatively slow. This is reasonable, as we do not divide the expected sample sizes by the optimal performance in each case, but with a strict lower bound on it instead.

Fifth, comparing Figure 2.2a with Figure 2.2c and 2.2e, we verify that knowledge of the exact number of signals roughly halves the required expected sample size in comparison to the case that we only have a lower and an upper bound on the number of signals.

Finally, we see by Tables 2.1a and 2.1b that the upper bounds (2.5) and (2.13) on the error probabilities are very crude. Nevertheless, from Figure 2.3a and 2.3b, we observe that using these conservative thresholds in the design of the proposed procedures leads to bounded performance loss as the error probabilities go to 0 relative to the case of sharp thresholds, obtained via Monte Carlo simulation. This is expected, as the

expected sample size scales with the logarithm of the error probabilities.

m	$\mathsf{P}_{\mathcal{A}}(d_G \neq \mathcal{A})$	$\mathsf{E}_{\mathcal{A}}(T_G)$	Upper bound
1	5.041E-05 (3.101E-07)	64.071 (0.157)	4.086E-4
3	6.034E-05 (5.343E-07)	78.386 (0.157)	9.534E-4
5	6.145E-05 (5.859E-07)	81.070 (0.156)	1.135E-3

(a) $\mathcal{P} = \mathcal{P}_m$. (T_G, d_G) with $c = 10$.

$\lvert A\rvert$	$\mathsf{P}_{\mathcal{A}}(\mathcal{A} \lesssim d_{GI})$	$\mathsf{E}_{\mathcal{A}}(T_{GI})$	Upper bound
3	3.653E-05 (5.447E-07)	142.173 (0.264)	4.540E-04
4	3.144E-05 (2.189E-07)	152.873 (0.264)	4.281E-04
5	2.621E-05 (1.825E-07)	152.895 (0.263)	3.891E-04
7	3.104E-07 (1.340E-08)	142.363 (0.270)	2.724E-04

(b) $\mathcal{P} = \mathcal{P}_{3,7}$. (T_{GI}, d_{GI}) with $b = 10$.

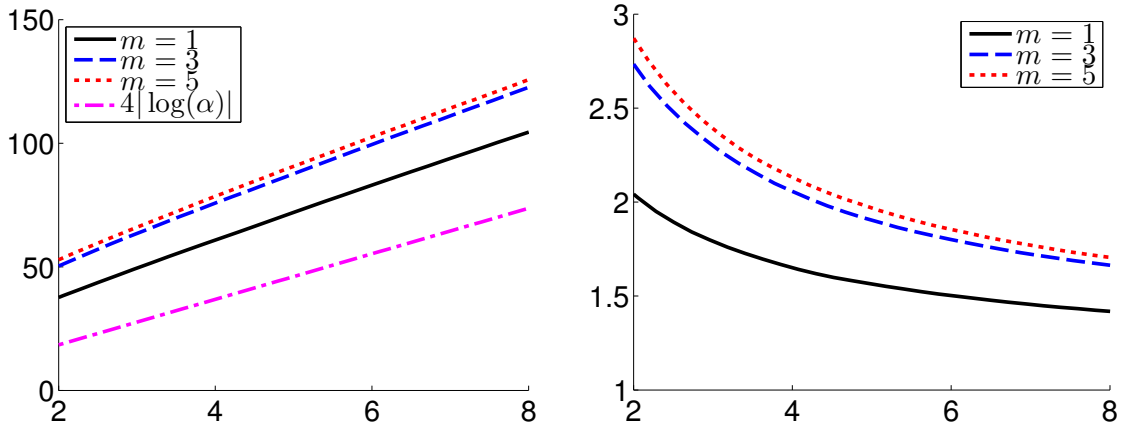
Table 2.1: The standard error of the estimate is included in the parenthesis. The upper bound is on the error control given by (2.5) for the first table and by (2.13) for the second.

2.7 Conclusions

We considered the problem of simultaneously testing multiple simple null hypotheses, each of them against a simple alternative, in a sequential setup. That is, the data for each testing problem are acquired sequentially and the goal is to stop sampling as soon as possible, simultaneously in all streams, and make a correct decision for each individual testing problem. The main goal of this Chapter was to propose feasible, yet asymptotically optimal, procedures that incorporate prior information on the number of signals (correct alternatives), and also to understand the potential gains in efficiency by such prior information.

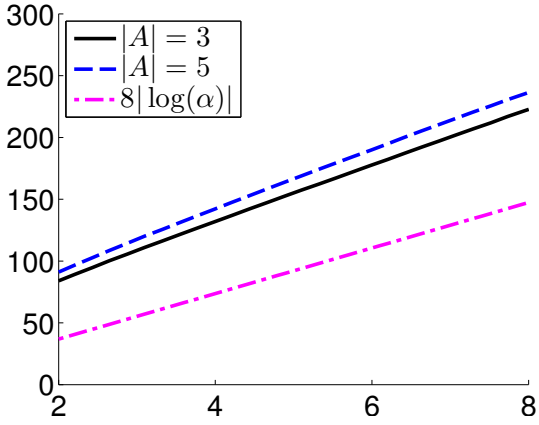
We studied this problem under the assumption that the data streams for the various hypotheses are independent. Without any distributional assumptions on the data that are acquired in each stream, we proposed procedures that control the probabilities of at least one false positive and at least one false negative below arbitrary user-specified levels. This was achieved in two general cases regarding the available prior information: when the exact number of signals is known in advance, and when we only have an upper and a lower bound for it. Furthermore, we proposed a Monte Carlo simulation method, based on importance sampling, that can facilitate the specification of non-conservative critical values for the proposed multiple testing procedures in practice. More importantly, in the special case of i.i.d. data in each stream, we were able to show that the proposed multiple testing procedures are asymptotically optimal, in the sense that they require the minimum possible expected sample size to a first-order asymptotic approximation as the error probabilities vanish at arbitrary rates.

These asymptotic optimality results have some interesting ramifications. First of all, they imply that any

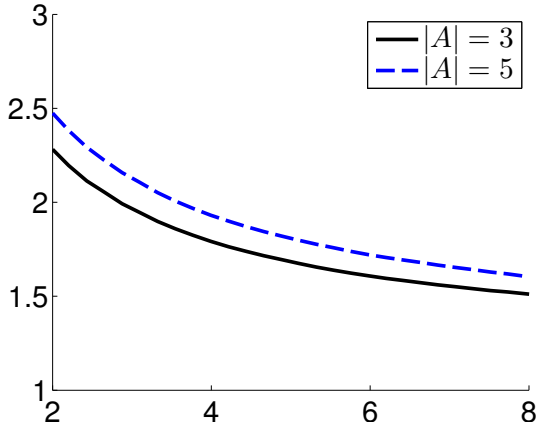


(a) Gap rule: $\mathcal{P} = \mathcal{P}_m$, $m = 1, 3, 5$

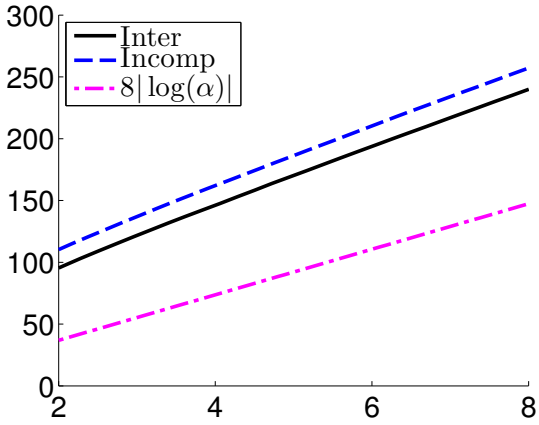
(b) Normalized by $4|\log(\alpha)|$



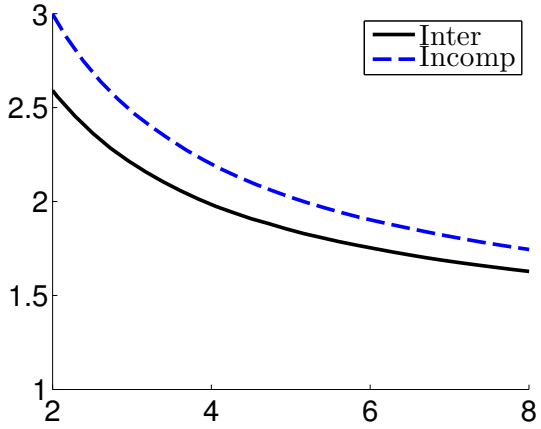
(c) Gap-intersection rule: $\mathcal{P} = \mathcal{P}_{3,7}$



(d) Normalized by $8|\log(\alpha)|$

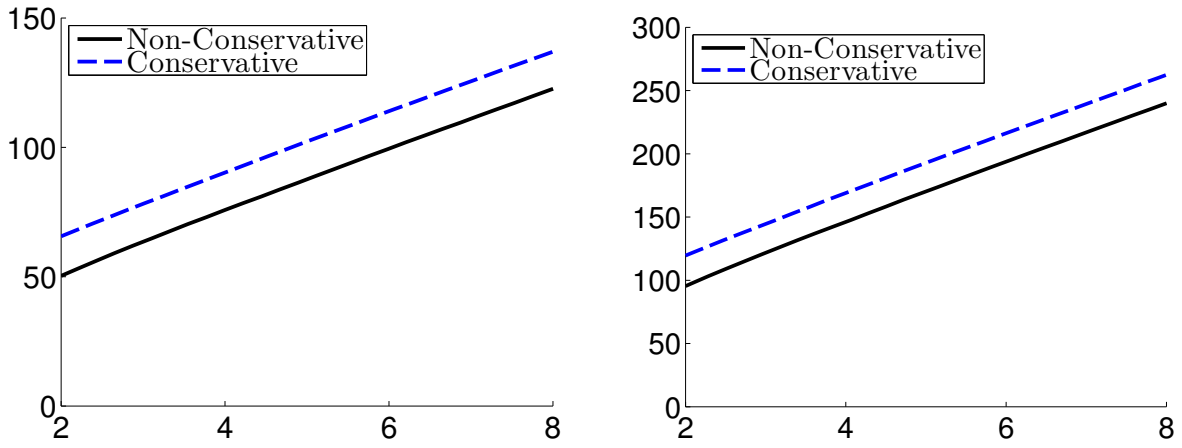


(e) Intersection vs Incomplete: $\mathcal{P} = \mathcal{P}_{0,10}$



(f) Normalized by $8|\log(\alpha)|$

Figure 2.2: The x-axis in all graphs is $|\log_{10}(\alpha)|$. In the first column, the y-axis denotes the expected sample size under $\mathsf{P}_{\mathcal{A}}$ that is required in order to control the *maximal* familywise type I error probability *exactly* at level α. The dash-dot lines in each plot correspond to the first-order approximation, which is also a lower bound, to the optimal expected sample size for the class $\Delta_{\alpha,\alpha}(\mathcal{P})$; due to symmetry, this lower bound does not depend on $|\mathcal{A}|$ in each setup. In the second column, we normalize each curve by its corresponding lower bound.



(a) Gap rule: $\mathcal{P} = \mathcal{P}_3$ (b) Intersection rule: $\mathcal{P} = \mathcal{P}_{0,10}$

Figure 2.3: The x-axis is $|\log_{10}(\alpha)|$, where α is user-specified level. The y-axis is the expected sample size. The dashed line uses the upper bound on the error probability to get conservative critical value, while the solid line uses the Monte Carlo approach to determine non-conservative threshold such that the *maximal* familywise type I error is controlled *exactly* at level α.

refinements of the proposed procedures, for example using a more judicious choice of alpha-spending and beta-spending functions, cannot reduce the expected sample size *to a first-order* asymptotic approximation. Second, they imply that bounds on the number of signals do not improve the minimum possible expected sample size *to a first-order asymptotic approximation*, apart from a very special case. On the other hand, knowledge of the *exact* number of signals does reduce the minimum possible expected sample size to a first order approximation, roughly by a factor of 2. These insights were corroborated by a simulation study, which however also revealed the limitations of a first-order asymptotic analysis and emphasized the importance of second-order terms.

To our knowledge, these are the first results on the asymptotic optimality of multiple testing procedures, with or without prior information, that control the familywise error probabilities of both types. However, there are still some important open questions that remain to be addressed. Do the proposed procedures attain, in the i.i.d. setup, the optimal expected sample size to a *second-order* asymptotic approximation as well? Does the first-order asymptotic optimality property remain valid for more general, non-i.i.d. data in the streams? While we conjecture that the answer to both these questions is affirmative, we believe that the corresponding proofs require different techniques from the ones we have used in the current Chapter.

There are also interesting generalizations of the setup we considered in this Chapter. For example, it is interesting to consider the sequential multiple testing problem when the goal is to control generalized error rates, such as the false discovery rate [6], instead of the more stringent familywise error rates. Another interesting direction is to allow the hypotheses in the streams to be specified up to an unknown parameter,

or to consider a non-parametric setup similarly to Li et al. [40]. Finally, it is still an open problem to design asymptotically optimal multiple testing procedures that incorporate prior information on the number of signals when it is possible and desirable to stop sampling at different times in the various streams.

2.8 Two lemmas

2.8.1 An information-theoretic inequality

In the proof of Theorem 2.3 we use the following, well-known, information-theoretic inequality, whose proof can be found, e.g., in Tartakovsky et al. [71] (Chapter 3.2).

Lemma 2.3. *Let* Q, P *be equivalent probability measures on a measurable space* $(\Omega, \mathcal{G})$ *and recall the function* φ *defined in* (2.19)*. Then, for every* $A \in \mathcal{G}$ *we have*

$$\mathsf{E}_{\mathsf{Q}}\left[\log \frac{d\mathsf{Q}}{d\mathsf{P}}\right] \geq \varphi\left(Q(A), \mathsf{P}(A^c)\right).$$

2.8.2 A lemma on multiple random walks

For the proof of Lemmas 2.1 and 2.2 we need an upper bound on the expectation of the first time that multiple random walks, not necessarily independent, are simultaneously above given thresholds. We state here the corresponding result in some generality.

Thus, let $L \geq 2$ and suppose that for each $l \in [L]$ we have a sequence of i.i.d. random variables, $\{\xi_n^l, n \in \mathbb{N}\}$, such that $\mu_l = \mathsf{E}[\xi_1^l] > 0$ and $\mathrm{Var}[\xi_1^l] < \infty$. For each $l \in [L]$, let

$$S_n^l = \sum_{i=1}^{n} \xi_i^l, \quad n \in \mathbb{N}$$

be the corresponding random walk. Here, *no assumption is made on the dependence structure among these random walks.* For an arbitrary vector $(a_1, \ldots, a_L)$, consider the stopping time

$$T = \inf\left\{n \geq 1 \ : \ S_n^l \geq a_l \text{ for every } l \in [L]\right\}.$$

The following lemma provides an upper bound on the expected value of T. The proof is identical to the one in Theorem 2 in Mei [47]; thus we omit it. We stress that although the theorem in the reference assumes independent random walks, exactly the same proof applies to the case of dependent random walks.

Lemma 2.4. *As $a_1, \ldots, a_L \to \infty$,*

$$\mathsf{E}[T] \leq \max_{l \in [L]} \left(\frac{a_l}{\mu_l} \right) + O\left(\sum_{l \in [L]} \sqrt{\frac{a_l}{\mu_l}} \right) \leq \max_{l \in [L]} \left(\frac{a_l}{\mu_l} \right) + O\left(L \sqrt{\max_{l \in [L]} \{a_l\}} \right).$$

Chapter 3

Sequential multiple testing with generalized error metrics

3.1 Introduction

[1] In the early development of multiple testing, the focus was on procedures that control the probability of at least *one* false positive, i.e., falsely rejected null [28, 29, 46]. As this requirement can be prohibitive when the number of hypotheses is large, the emphasis gradually shifted to the control of less stringent error metrics, such as (i) the expectation [8] or the quantiles [36] of the *false discovery proportion*, i.e., the proportion of false positives among the rejected nulls, and (ii) the *generalized familywise error rate*, i.e., the probability of at least $k \geq 1$ false positives [30, 36]. During the last two decades, various procedures have been proposed to control the above error metrics [9, 27, 59, 60]. Further, the problem of maximizing the number of true positives subject to a generalized control on false positives has been studied in [37, 55, 69, 70], whereas in [14] the false negatives are incorporated into the risk function in a Bayesian decision theoretic framework.

In this Chapter, we consider the same sequential multiple testing setup as in Chapter 2, but instead focus on two related, yet distinct, generalized error metrics. The first one is a generalization of the usual mis-classification rate [39, 45], where the probability of at least $k \geq 1$ mistakes, of any kind, is controlled. The second one controls generalized familywise error rates of both types [3, 19], i.e., the probabilities of at least $k_1 \geq 1$ false positives and at least $k_2 \geq 1$ false negatives.

Various sequential procedures have been proposed recently to control such generalized familywise error rates [3, 5, 7, 17, 18, 19]. To the best of our knowledge, the efficiency of these procedures is understood only in the case of *classical* familywise error rates, i.e., when $k_1 = k_2 = 1$. Specifically, in the case of independent streams with i.i.d. observations, an asymptotic lower bound was obtained in [66] for the optimal expected sample size (ESS) as the error probabilities go to 0, and was shown to be attained, under any signal configuration, by several existing procedures. However, the results in [66] do not extend to *generalized* error metrics, since the technique for the proof of the asymptotic lower bound requires that the probability of not identifying the correct subset of signals goes to 0. Further, as we shall see, existing procedures fail to be

[1] This chapter is based on my publication [67].

asymptotically optimal in general under generalized error metrics.

The lack of an optimality theory under such generalized error control implies that it is not well understood how the best possible ESS depends on the user-specified parameters. This limits the applicability of generalized error metrics, as it is not clear for the practitioner how to select the number of hypotheses to be "sacrificed" for the sake of a faster decision.

In this Chapter, we address this research gap by developing an asymptotic optimality theory for the sequential multiple testing problem under the two generalized error metrics mentioned above. Specifically, for each formulation we characterize the optimal ESS as the error probabilities go to 0, and propose a novel, feasible sequential multiple testing procedure that achieves the optimal ESS under every signal configuration. These results are established under the assumption of independent data streams, and require that the log-likelihood ratio statistic in each stream satisfies a certain Strong Law of Large Numbers. Thus, even in the case of classical familywise error rates, we extend the corresponding results in [66] by relaxing the i.i.d. assumption in each stream.

Finally, whenever sequential testing procedures are utilized, it is of interest to quantify the savings in the ESS over fixed-sample size schemes with the same error control guarantees. In the case of i.i.d. data streams, we obtain an asymptotic lower bound for the gains of sequential sampling over *any* fixed-sample size schemes, and also characterize the asymptotic gains over a specific fixed-sample size procedure.

In order to convey the main ideas and results with the maximum clarity, we first consider the case that the local hypotheses are simple, and then extend our results to the case of composite hypotheses. Thus, the remainder of the Chapter is organized as follows: in Section 3.2 we formulate the two problems of interest in the case of simple hypotheses. The case of generalized mis-classification rate is presented in Section 3.3, and the case of generalized familywise error rates in Section 3.4. In Section 3.5 we present two simulation studies under the second error metric. In Section 3.6 we extend our results to the case of composite hypotheses. We conclude and discuss potential extensions of this Chapter in Section 3.7. From Section 3.8-3.14, we present proofs, more simulation studies and a detailed analysis of the case of composite hypotheses. For convenience, we list in Table 3.1 the procedures that are considered in this Chapter.

3.2 Problem formulation

Consider *independent* streams of observations, $X^j := \{X^j(n) : n \in \mathbb{N}\}$, where $j \in [J] := \{1, \ldots, J\}$ and $\mathbb{N} := \{1, 2, \ldots\}$. For each $j \in [J]$, we denote by P^j the distribution of X^j and consider two simple hypotheses

Procedure	Metric	Section	Main results	Conditions for AO
Sum-Intersection†	GMIS	3.3.1	Thrm 3.3	(3.8)
Leap†	GFWER	3.4.2	Thrm 3.6	(3.8)
Asym. Sum-Intersection†	GFWER	3.4.1	Cor 3.2	(3.8) + (3.11) + (3.12)
Intersection	both	3.2.2	Cor 3.1/ 3.2	(3.8) + (3.11) / (3.12)
MNP (fixed sample)	both	3.2.3	Thrm 3.4/ 3.7	Not optimal

Table 3.1: Procedures marked with † are novel. Procedures in bold font are asymptotically optimal (AO) without requiring special structure. GMIS is short for generalized mis-classification rate, and GFWER for generalized familywise error rates.

for it,

$$\mathsf{H}_0^{\mathsf{j}} : \mathsf{P}^{\mathsf{j}} = \mathsf{P}_0^{\mathsf{j}} \text{ versus } \mathsf{H}_1^{\mathsf{j}} : \mathsf{P}^{\mathsf{j}} = \mathsf{P}_1^{\mathsf{j}}. \tag{3.1}$$

We denote by P_A the distribution of $(X^1, \ldots, X^J)$ when $A \subset [J]$ is the subset of data streams with signal, i.e., in which the alternative hypothesis is correct. Due to the assumption of independence among streams, P_A is the following product measure:

$$\mathsf{P}_A := \bigotimes_{j=1}^{J} \mathsf{P}^j; \quad \mathsf{P}^j = \begin{cases} \mathsf{P}_0^j, & \text{if } j \notin A \\ \mathsf{P}_1^j, & \text{if } j \in A. \end{cases} \tag{3.2}$$

Moreover, we denote by $\mathcal{F}_n^j$ the σ-field generated by the first n observations in the j-th stream, i.e., $\sigma(X^j(1), \ldots, X^j(n))$, and by $\mathcal{F}_n$ the σ-field generated by the first n observations in all streams, i.e., $\sigma(\mathcal{F}_n^j,\, j \in [J])$, where $n \in \mathbb{N}$.

Assuming that the data in all streams become available *sequentially*, the goal is to stop sampling *as soon as possible*, and upon stopping to solve the J hypothesis testing problems subject to certain error control guarantees. Formally, a *sequential multiple testing procedure* is a pair $\delta = (T, D)$ where T is an $\{\mathcal{F}_n\}$-stopping time at which sampling is terminated in all streams, and D an $\mathcal{F}_T$-measurable, J-dimensional vector of Bernoullis, $(D^1, \ldots, D^J)$, so that the alternative hypothesis is selected in the j-th stream if and only if $D^j = 1$. With an abuse of notation, we also identify D with the rejected nulls, i.e., the subset of streams in which the alternative hypothesis is selected upon stopping, i.e., $\{j \in [J] :\ D^j = 1\}$.

We consider two kinds of error control, which lead to two different problems. Their main difference is that the first one does not differentiate between *false positives*, i.e., rejecting the null when it is correct, and *false negatives*, i.e., accepting the null when it is false. Specifically, in the first one we control the generalized mis-classification rate, i.e., the probability of committing *at least k mistakes, of any kind,* where k is a user-specified integer such that $1 \leq k < J$. When A is the true subset of signals, a decision rule D makes at

least k mistakes, of any kind, if D and A differ in at least k components, i.e., $|A \triangle D| \geq k$, where for any two sets A and D, $A \triangle D$ is their symmetric difference, i.e. $(A \setminus D) \cup (D \setminus A)$, and $|\cdot|$ denotes set-cardinality. Thus, given tolerance level $\alpha \in (0,1)$, the class of multiple testing procedures of interest in this case is

$$\Delta_k(\alpha) := \left\{ (T, D) : \max_{A \subset [J]} \mathsf{P}_A(|A \triangle D| \geq k) \leq \alpha \right\}.$$

Then, the first problem is formulated as follows:

Problem 3.1. *Given a user-specified integer k in $[1, J)$, find a sequential multiple testing procedure that (i) controls the generalized mis-classification rate, i.e., it can be designed to belong to $\Delta_k(\alpha)$ for any given α, and (ii) achieves the smallest possible expected sample size,*

$$N_A^*(k, \alpha) := \inf_{(T,D) \in \Delta_k(\alpha)} \mathsf{E}_A[T],$$

for every $A \subset [J]$, to a first-order asymptotic approximation as $\alpha \to 0$.

In the second problem of interest in this work, we control generalized familywise error rates of both types, i.e., the probabilities of *at least k_1 false positives and at least k_2 false negatives*, where $k_1, k_2 \geq 1$ are integers such that $k_1 + k_2 \leq J$. When the true subset of signals is A, a decision rule D makes at least k_1 false positives when $|D \setminus A| \geq k_1$ and at least k_2 false negatives when $|A \setminus D| \geq k_2$. Thus, given tolerance levels $\alpha, \beta \in (0,1)$, the class of procedures of interest in this case is

$$\begin{aligned} \Delta_{k_1,k_2}(\alpha, \beta) := \{(T, D) : & \max_{A \subset [J]} \mathsf{P}_A(|D \setminus A| \geq k_1) \leq \alpha \quad \text{and} \\ & \max_{A \subset [J]} \mathsf{P}_A(|A \setminus D| \geq k_2) \leq \beta\}. \end{aligned} \tag{3.3}$$

Then, the second problem is formulated as follows:

Problem 3.2. *Given user-specified integers $k_1, k_2 \geq 1$ such that $k_1 + k_2 \leq J$, find a sequential multiple testing procedure that (i) simultaneously controls generalized familywise error rates of both types, i.e., it can be designed to belong to $\Delta_{k_1,k_2}(\alpha, \beta)$ for any given $\alpha, \beta \in (0,1)$, and (ii) achieves the smallest possible expected sample size,*

$$N_A^*(k_1, k_2, \alpha, \beta) := \inf_{(T,D) \in \Delta_{k_1,k_2}(\alpha,\beta)} \mathsf{E}_A[T],$$

for every $A \subset [J]$, to a first-order asymptotic approximation as α and β go to 0, at arbitrary rates.

3.2.1 Assumptions

We now state the assumptions that we will make in the next two sections in order to solve these two problems. First of all, for each $j \in [J]$ we assume that the probability measures P_0^j and P_1^j in (3.1) are mutually absolutely continuous when restricted to $\mathcal{F}_n^j$, and we denote the corresponding log-likelihood ratio (LLR) statistic as follows:

$$\lambda^j(n) := \log \frac{d\mathsf{P}_1^j}{d\mathsf{P}_0^j}(\mathcal{F}_n^j), \quad \text{for } n \in \mathbb{N}.$$

For $A, C \subset [J]$ and $n \in \mathbb{N}$ we denote by $\lambda^{A,C}(n)$ the LLR of P_A versus P_C when both measures are restricted to $\mathcal{F}_n$, and from (3.2) it follows that

$$\lambda^{A,C}(n) := \log \frac{d\mathsf{P}_A}{d\mathsf{P}_C}(\mathcal{F}_n) = \sum_{j \in A \setminus C} \lambda^j(n) - \sum_{j \in C \setminus A} \lambda^j(n). \tag{3.4}$$

In order to guarantee that the proposed multiple testing procedures terminate almost surely and satisfy the desired error control, it will suffice to assume that

$$\mathsf{P}_1^j \left(\lim_{n \to \infty} \lambda^j(n) = \infty \right) = \mathsf{P}_0^j \left(\lim_{n \to \infty} \lambda^j(n) = -\infty \right) = 1 \quad \forall\, j \in [J]. \tag{3.5}$$

In order to establish an asymptotic lower bound on the optimal ESS for each problem, we will need the stronger assumption that for each $j \in [J]$ there are positive numbers, $\mathcal{I}_1^j, \mathcal{I}_0^j$, such that the following Strong Law of Large Numbers (SLLN) hold:

$$\mathsf{P}_1^j \left(\lim_{n \to \infty} \frac{\lambda^j(n)}{n} = \mathcal{I}_1^j \right) \;=\; \mathsf{P}_0^j \left(\lim_{n \to \infty} \frac{\lambda^j(n)}{n} = -\mathcal{I}_0^j \right) \;=\; 1. \tag{3.6}$$

When the LLR statistic in each stream has *independent and identically distributed (i.i.d.) increments*, the SLLN (3.6) will also be sufficient for establishing the asymptotic optimality of the proposed procedures. When this is not the case, we will need an assumption on the rate of convergence in the SLLN (3.6). Specifically, we will need to assume that for every $\epsilon > 0$ and $j \in [J]$,

$$\sum_{n=1}^{\infty} \mathsf{P}_1^j \left(\left| \frac{\lambda^j(n)}{n} - \mathcal{I}_1^j \right| > \epsilon \right) < \infty, \quad \sum_{n=1}^{\infty} \mathsf{P}_0^j \left(\left| \frac{\lambda^j(n)}{n} + \mathcal{I}_0^j \right| > \epsilon \right) < \infty. \tag{3.7}$$

Condition (3.7) is known as *complete* convergence [31], and is a stronger assumption than (3.6), due to the Borel-Cantelli lemma. This condition is satisfied in various testing problems where the observations in each data stream are dependent, such as autoregressive time-series models and state-space models. For more

details, we refer to [71, Chapter 3.4].

To sum up, the only distributional assumption for our asymptotic optimality theory is that the LLR statistic in each stream

$$\begin{aligned} &\text{either has i.i.d. increments and satisfies the SLLN (3.6),} \\ &\text{or satisfies the SLLN with complete convergence (3.7).} \end{aligned} \tag{3.8}$$

Remark 3.1. *If* (3.6) *(resp.* (3.7)*) holds, the normalized LLR,* $\lambda^{A,C}(n)/n$*, defined in* (3.4)*, converges almost surely (resp. completely) under* P_A *to*

$$\mathcal{I}^{A,C} := \sum_{i \in A \setminus C} \mathcal{I}_1^i + \sum_{j \in C \setminus A} \mathcal{I}_0^j. \tag{3.9}$$

The numbers $\mathcal{I}^{A,C}$ *and* $\mathcal{I}^{C,A}$ *will turn out to determine the inherent difficulty in distinguishing between* P_A *and* P_C *and will play an important role in characterizing the optimal performance under* P_A *and* P_C *respectively.*

3.2.2 The Intersection rule

To the best of our knowledge, Problem 3.2 has been solved only under the assumption of i.i.d. data streams and *only in the case of classical error control, that is when* $k_1 = k_2 = 1$ [66]. An asymptotically optimal procedure in this setup is the so-called "*Intersection*" rule, $\delta_I := (T_I, D_I)$, proposed in [17, 18], where

$$\begin{aligned} T_I &:= \inf\left\{n \geq 1 : \lambda^j(n) \notin (-a, b) \text{ for every } j \in [J]\right\}, \\ D_I &:= \left\{j \in [J] : \lambda^j(T_I) > 0\right\}, \end{aligned} \tag{3.10}$$

and a, b are positive thresholds. This procedure requires the local test statistic in *every* stream to provide sufficiently strong evidence for the sampling to be terminated. The Intersection rule was also shown in [19] to control *generalized* familywise error rates, however its efficiency in this setup remains an open problem, even in the case of i.i.d. data streams. Our asymptotic optimality theory in the next sections will reveal that the Intersection rule is asymptotically optimal with respect to Problems 3.1 and 3.2 only when the multiple testing problem satisfies *a very special structure.*

Definition 3.1. *We say that the multiple testing problem is*

(i) symmetric, if for every $j \in [J]$ *the distribution of* λ^j *under* P_0^j *is the same as the distribution of* $-\lambda^j$ *under* P_1^j*,*

(ii) homogeneous, if for every $j \in [J]$ the distribution of λ^j under P_i^j does not depend on j, where $i \in \{0,1\}$.

It is clear that when the multiple testing problem is both *symmetric and homogeneous*, we have

$$\mathcal{I}_0^j = \mathcal{I}_1^j = \mathcal{I} \quad \text{for every } j \in [J]. \tag{3.11}$$

In the next sections we will show that the Intersection rule is asymptotically optimal for Problem 3.1 when (3.11) holds, whereas its asymptotic optimality with respect to Problem 3.2 will *additionally* require that the user-specified parameters satisfy the following conditions:

$$k_1 = k_2 \quad \text{and} \quad \alpha = \beta. \tag{3.12}$$

3.2.3 Fixed-sample size schemes

Let $\Delta_{fix}(n)$ denote the class of procedures for which the decision rule depends on the data collected up to a *deterministic* time n, i.e.,

$$\Delta_{fix}(n) := \{(n, D) : \ D \subset [J] \ \text{ is } \mathcal{F}_n\text{-measurable}\}.$$

For any given integers $k, k_1, k_2 \geq 1$ with $k, k_1 + k_2 < J$ and $\alpha, \beta \in (0,1)$, let

$$\begin{aligned} n^*(k,\alpha) &:= \inf\left\{n \in \mathbb{N} : \ \Delta_{fix}(n) \bigcap \Delta_k(\alpha) \neq \emptyset\right\}, \\ n^*(k_1,k_2,\alpha,\beta) &:= \inf\left\{n \in \mathbb{N} : \ \Delta_{fix}(n) \bigcap \Delta_{k_1,k_2}(\alpha,\beta) \neq \emptyset\right\}, \end{aligned} \tag{3.13}$$

denote the minimum sample sizes required by *any fixed-sample size scheme* under the two error metrics of interest. In the case of i.i.d. observations in the data streams, we establish *asymptotic lower bounds* for the above two quantities as the error probabilities go to 0. To the best of our knowledge, there is no fixed-sample size procedure that attains these bounds. For this reason, we also study a specific procedure that runs a *Neyman-Pearson test at each stream*. Formally, this procedure is defined as follows:

$$\delta_{NP}(n,h) := (n, D_{NP}(n,h)), \ \ D_{NP}(n,h) := \{j \in [J] : \lambda^j(n) > nh_j\}, \tag{3.14}$$

where $h = (h_1, \ldots, h_J) \in \mathbb{R}^J$ and $n \in \mathbb{N}$, and we will we refer to it as *multiple Neyman-Pearson (MNP) rule*. In the case of generalized mis-classification rate, we characterize the minimum sample size required by this

procedure,

$$n_{NP}(k,\alpha) := \inf\{n \in \mathbb{N} : \ \exists\, h \in \mathbb{R}^J, \ \ \delta_{NP}(n,h) \in \Delta_k(\alpha)\},$$

to a first-order approximation as $\alpha \to 0$. In the case of generalized familywise error rates, for simplicity of presentation we further restrict ourselves to *homogeneous*, but not necessarily symmetric, multiple testing problems, and characterize the asymptotic minimum sample size required by the MNP rule that utilizes the same threshold in each stream, i.e.,

$$\hat{n}_{NP}(k_1,k_2,\alpha,\beta) := \inf\{n \in \mathbb{N} : \exists\, h \in \mathbb{R}, \ \ \delta_{NP}(n,h\mathbf{1}_J) \in \Delta_{k_1,k_2}(\alpha,\beta)\},$$

where $\mathbf{1}_J \in \mathbb{R}^J$ is a J-dimensional vector of ones.

3.2.4 The i.i.d. case

As mentioned earlier, our asymptotic optimality theory will apply whenever condition (3.8) holds, thus, beyond the case of i.i.d. data streams. However, our analysis of *fixed-sample size* schemes will rely on large deviation theory [20] and will be focused on the i.i.d. case. Thus, it is convenient to introduce some relevant notations for this setup.

Specifically, when for each $j \in [J]$ the observations in the j-th stream are independent with common density f^j relative to a σ-finite measure ν^j, the hypothesis testing problem (3.1) takes the form

$$\mathsf{H}_0^j \ : f^j = f_0^j \ \text{ versus } \ \mathsf{H}_1^j \ : f^j = f_1^j, \tag{3.15}$$

and $\mathcal{I}_1^j, \mathcal{I}_0^j$ correspond to the *Kullback-Leibler divergences* between f_1^j and f_0^j, i.e.,

$$\mathcal{I}_1^j = \int \log\left(f_1^j/f_0^j\right) f_1^j \, d\nu^j, \quad \mathcal{I}_0^j = \int \log\left(f_0^j/f_1^j\right) f_0^j \, d\nu^j. \tag{3.16}$$

In this case, each LLR statistic λ^j has i.i.d. increments, and (3.8) is satisfied as long as $\mathcal{I}_1^j$ and $\mathcal{I}_0^j$ are both positive and finite. For each $j \in [J]$, we further introduce the convex conjugate of the cumulant generating function of $\lambda^j(1)$

$$z \in \mathbb{R} \to \Phi^j(z) := \sup_{\theta \in \mathbb{R}} \left\{z\theta - \Psi^j(\theta)\right\}, \text{ where } \Psi^j(\theta) := \log \mathsf{E}_0^j\left[e^{\theta\lambda^j(1)}\right]. \tag{3.17}$$

The value of Φ^j at zero is the *Chernoff information* [20] for the testing problem (3.15), and we will denote

it as $\mathcal{C}^j$, i.e., $\mathcal{C}^j := \Phi^j(0)$.

Finally, we will illustrate our general results in the case of testing normal means. Hereafter, $\mathcal{N}$ denotes the density of the normal distribution.

Example 3.1. *If $f_0^j = \mathcal{N}(0, \sigma_j^2)$ and $f_1^j = \mathcal{N}(\mu_j, \sigma_j^2)$ for all $j \in [J]$, then*

$$\lambda^j(1) = \theta_j^2 \left(X^j(1)/\mu_j - 1/2 \right), \quad \textit{where} \quad \theta_j := \mu_j/\sigma_j.$$

Consequently the multiple testing problem is symmetric and

$$\mathcal{I}^j := \mathcal{I}_0^j = \mathcal{I}_1^j = \theta_j^2/2, \quad \Phi^j(z) = (z + \mathcal{I}^j)^2/(4\mathcal{I}^j) \text{ for any } z \in \mathbb{R}. \tag{3.18}$$

3.2.5 Notation

Finally, we collect some notations that will be used extensively throughout the Chapter: C_k^J denotes the binomial coefficient $\binom{J}{k}$, i.e., the number of subsets of size k from a set of size J; $a \vee b$ represents $\max\{a, b\}$; $x \sim y$ means that $\lim_y x/y = 1$ and $x(b) = o(1)$ that $\lim_b x(b) = 0$. $\mathbb{N} := \{1, 2, \ldots\}$, $[J] := \{1, \ldots, J\}$. For any two sets A, B, $A \triangle B$ is the symmetric difference, $(A \setminus B) \cup (B \setminus A)$, and $|\cdot|$ denotes set-cardinality.

3.3 Generalized mis-classification rate

In this section we consider Problem 3.1, and carry out the following program: first, we propose a novel procedure that controls the generalized mis-classification rate. Then, we establish an asymptotic lower bound on the optimal ESS and show that it is attained by the proposed scheme. As a corollary, we show that the Intersection rule is asymptotically optimal when (3.11) holds. Finally, we make a comparison with fixed-sample size procedures in the i.i.d. case (3.15).

3.3.1 Sum-Intersection rule

In order to implement the proposed procedure, which we will denote $\delta_S(b) := (T_S(b), D_S(b))$, we need at each time $n \in \mathbb{N}$ prior to stopping to order the *absolute values* of the local LLR statistics, $|\lambda^j(n)|, j \in [J]$. If we denote the corresponding ordered values by

$$\widetilde{\lambda}^1(n) \leq \ldots \leq \widetilde{\lambda}^J(n),$$

we can think of $\widetilde{\lambda}^1(n)$ (resp. $\widetilde{\lambda}^J(n)$) as the least (resp. most) "significant" local test statistic at time n, in the sense that it provides the weakest (resp. strongest) evidence in favor of either the null or the alternative. Then, sampling is terminated at the first time the *sum of the k least significant local LLRs* exceeds some positive threshold b, and the null hypothesis is rejected in every stream that has a positive LLR upon stopping, i.e.,

$$T_S(b) := \inf\left\{ n \geq 1 : \sum_{j=1}^{k} \widetilde{\lambda}^j(n) \geq b \right\}, \; D_S(b) := \left\{ j \in [J] : \; \lambda^j(T_S(b)) > 0 \right\}.$$

The threshold b is selected to guarantee the desired error control. When $k = 1$, $\delta_S(b)$ coincides with the Intersection rule, $\delta_I(b, b)$, defined in (3.10). When $k > 1$, the two rules are different but they share a similar flavor, since $\delta_S(b)$ stops the first time *all sums of the form* $\sum_{j \in A} |\lambda^j(n)|$, *with* $A \subset [J]$ *and* $|A| = k$, are simultaneously above b. For this reason, we refer to $\delta_S(b)$ as *Sum-Intersection rule*. Hereafter, we will typically suppress the dependence of $\delta_S(b)$ on threshold b in order to lighten the notation.

3.3.2 Error control of the Sum-Intersection rule

For any choice of threshold b, the Sum-Intersection rule clearly terminates almost surely, under every signal configuration, as long as condition (3.5) holds. In the next theorem we show how to select b to guarantee the desired error control. We stress that no additional distributional assumptions are needed for this purpose.

Theorem 3.1. *Assume* (3.5) *holds. For any* $\alpha \in (0, 1)$ *we have* $\delta_S(b_\alpha) \in \Delta_k(\alpha)$ *when*

$$b_\alpha = |\log(\alpha)| + \log(C_k^J). \tag{3.19}$$

Proof. The proof can be found in Section 3.9.1. □

The choice of b suggested by the previous theorem will be sufficient for establishing the asymptotic optimality of the Sum-Intersection rule, but may be conservative for practical purposes. In the absence of more accurate approximations for the error probabilities, we recommend finding the value of b for which the target level is attained using Monte Carlo simulation. This means simulating off-line, i.e., before the sampling process begins, for every $A \subset [J]$ the error probability $\mathsf{P}_A(|A \triangle D_S(b)| \geq k)$ for various values of b, and then selecting the value for which the maximum of these probabilities over $A \subset [J]$ matches the nominal level α.

This simulation task is significantly facilitated when the multiple testing problem has a special structure. If the problem is *symmetric*, for any given threshold b the error probabilities coincide for all $A \subset [J]$, and thus

it suffices to simulate the error probability under a single measure, e.g., $\mathsf{P}_\emptyset$. If the problem is *homogeneous*, the error probabilities depend only on the size of A, not the actual subset. Thus, it suffices to simulate the above probabilities for at most $(J+1)$ configurations. Similar ideas apply in the presence of block-wise homogeneity.

Moreover, it is worth pointing out that when b is large, we can apply importance sampling techniques to simulate the corresponding "small" error probabilities, similarly to [65].

3.3.3 Asymptotic lower bound on the optimal performance

We now obtain an asymptotic (as $\alpha \to 0$) lower bound on $N_A^*(k,\alpha)$, the optimal ESS when the true subset of signals is A, for any given $k \geq 1$. When $k = 1$, from [72, Theorem 2.2] it follows that such a lower bound is given by $|\log(\alpha)|/\min_{C \neq A} \mathcal{I}^{A,C}$, where $\mathcal{I}^{A,C}$ is defined in (3.9). Thus, the asymptotic lower bound when $k = 1$ is determined by the "wrong" subset that is the most difficult to be distinguished from A, where the difficulty level is measured by the information numbers defined in (3.9).

The techniques in [72] require that the probability of selecting the wrong subset goes to 0, thus, they do not apply to the case of generalized error control ($k > 1$). Nevertheless, it is reasonable to conjecture that the corresponding asymptotic lower bound when $k > 1$ will still be determined by the wrong subset that is the most difficult to be distinguished from A, with the difference that a subset will now be "wrong" under P_A *if it differs from* A *in at least* k *components*, i.e., if it does *not* belong to

$$\mathcal{U}_k(A) := \{C \subset [J] : |A \triangle C| < k\}.$$

This conjecture is verified by the following theorem.

Theorem 3.2. *If the SLLN* (3.6) *holds, then for any* $A \subset [J]$*, as* $\alpha \to 0$*,*

$$N_A^*(k,\alpha) \geq \frac{|\log(\alpha)|}{\mathcal{D}_A(k)}(1 - o(1)), \text{ where } \quad \mathcal{D}_A(k) := \min_{C \notin \mathcal{U}_k(A)} \mathcal{I}^{A,C}. \tag{3.20}$$

The proof in the case of the *classical* mis-classification rate ($k = 1$) is based on a change of measure from P_A to P_{A^*}, where A^* is chosen such that (i) A is a "wrong" subset under P_{A^*}, i.e., $A \neq A^*$ and (ii) A^* is "close" to A, in the sense that $\mathcal{I}^{A,A^*} \leq \mathcal{I}^{A,C}$ for every $C \neq A$ (see, e.g., [72, Theorem 2.2]).

When $k \geq 2$, there are more than one "correct" subsets under P_A. The key idea in our proof is that for *each* "correct" subset $B \in \mathcal{U}_k(A)$ we apply a different change of measure $\mathsf{P}_A \to \mathsf{P}_{B^*}$, where B^* is chosen such that (i) B is a "wrong" subset under P_{B^*}, i.e., $B \notin \mathcal{U}_k(B^*)$, and (ii) B^* is "close" to A, in the sense

that $I^{A,B^*} \leq \mathcal{I}^{A,C}$ for every $C \notin \mathcal{U}_k(A)$. The existence of such B^* is established in Section 3.9.2, and the proof of Theorem 3.2 is carried out in Section 3.9.3.

3.3.4 Asymptotic optimality

We are now ready to establish the asymptotic optimality of the Sum-Intersection rule by showing that it attains the asymptotic lower bound of Theorem 3.2 under every signal configuration.

Theorem 3.3. *Assume* (3.8) *holds. Then, for any $A \subset [J]$ we have as $b \to \infty$ that*

$$\mathsf{E}_A[T_S(b)] \leq \frac{b}{\mathcal{D}_A(k)}\,(1+o(1)). \tag{3.21}$$

When in particular b is selected such that $\delta_S \in \Delta_k(\alpha)$ and $b \sim |\log(\alpha)|$, e.g. as in (3.19)*, then for every $A \subset [J]$ we have as $\alpha \to 0$*

$$\mathsf{E}_A\,[T_S] \;\sim\; \frac{|\log\alpha|}{\mathcal{D}_A(k)} \;\sim\; N_A^*(k,\alpha).$$

Proof. If (3.21) holds and b is such that $\delta_S \in \Delta_k(\alpha)$ and $b \sim |\log(\alpha)|$, then δ_S attains the asymptotic lower bound in Theorem 3.2. Thus, it suffices to prove (3.21), which is done in the Section 3.9.4. □

The asymptotic characterization of the optimal ESS, $N_A^*(k,\alpha)$, illustrates the trade-off among the ESS, the number of mistakes to be tolerated, and the error tolerance level α. Specifically, it suggests that, for "small" values of α, tolerating $k-1$ mistakes reduces the ESS by a factor of $\mathcal{D}_A(k)/\mathcal{D}_A(1)$, which is *at least* k for every $A \subset [J]$. To justify the latter claim, note that if we denote the ordered information numbers $\{\mathcal{I}_1^j, j \in A\} \cup \{\mathcal{I}_0^j, j \notin A\}$ by $\widetilde{\mathcal{I}}^{(1)}(A) \leq \ldots \leq \widetilde{\mathcal{I}}^{(J)}(A)$, then

$$\mathcal{D}_A(k) \;=\; \sum_{j=1}^{k} \widetilde{\mathcal{I}}^{(j)}(A).$$

In the following corollary we show that the Intersection rule is asymptotically optimal when (3.11) holds, which is the case for example when the multiple testing problem is *both symmetric and homogeneous*.

Corollary 3.1. *(i) Assume* (3.5) *holds. For any $\alpha \in (0,1)$ we have $\delta_I(b,b) \in \Delta_k(\alpha)$ when b is equal to b_α/k, where b_α is defined in* (3.19)*.*
(ii) Suppose b is selected such that $\delta_I(b,b) \in \Delta_k(\alpha)$ and $b \sim |\log\alpha|/k$, e.g., as in (i). If (3.8) *holds, then*

$$\mathsf{E}_A\,[T_I] \leq \frac{|\log\alpha|}{k\mathcal{D}_A(1)}\,(1+o(1)).$$

If also (3.11) *holds, then for any* $A \subset [J]$ *we have as* $\alpha \to 0$ *that*

$$\mathsf{E}_A[T_I] \sim \frac{|\log \alpha|}{k\mathcal{I}} \sim N_A^*(k,\alpha).$$

Proof. The proof can be found in Section 3.9.5. □

Remark 3.2. *When* (3.11) *is violated, the Intersection rule fails to be asymptotically optimal. This will be illustrated with a simulation study in Section 3.8.2.*

3.3.5 Fixed-sample size rules

Finally, we focus on the i.i.d. case (3.15) and consider procedures that stop at a deterministic time, selected to control the generalized mis-classification rate. We recall that $\mathcal{C}^j$ is the Chernoff information in the j^{th} testing problem, and we denote by $\mathcal{B}(k)$ the sum of the smallest k local Chernoff informations, i.e.,

$$\mathcal{B}(k) := \sum_{j=1}^{k} \mathcal{C}^{(j)},$$

where $\mathcal{C}^{(1)} \leq \mathcal{C}^{(2)} \leq \ldots \leq \mathcal{C}^{(J)}$ are the ordered values of the local Chernoff information numbers $\mathcal{C}_j, j \in [J]$.

Theorem 3.4. *Consider the multiple testing problem with i.i.d. streams defined in* (3.15) *and suppose that the Kullback-Leibler numbers in* (3.16) *are positive and finite. For any user-specified integer* $1 \leq k \leq (J+1)/2$ *and* $A \subset [J]$*, we have as* $\alpha \to 0$

$$\frac{\mathcal{D}_A(k)}{\mathcal{B}(2k-1)}\,(1-o(1)) \;\leq\; \frac{n^*(k,\alpha)}{N_A^*(k,\alpha)} \;\leq\; \frac{n_{NP}(k,\alpha)}{N_A^*(k,\alpha)} \;\sim\; \frac{\mathcal{D}_A(k)}{\mathcal{B}(k)}.$$

Proof. The proof can be found in Section 3.9.6. □

Remark 3.3. *Since any fixed time is also a stopping time, the lower bound is relevant only when* $\mathcal{D}_A(k) > \mathcal{B}(2k-1)$ *for some* $A \subset [J]$.

We now specialize the results of the previous theorem to the *testing of normal means* (a Bernoulli example is presented in the Section 3.9.7). In Example 3.1 we saw that in the Gaussian case $\mathcal{C}^j = \mathcal{I}^j/4$ for every $j \in [J]$, which implies $\mathcal{D}_A(k) = 4\mathcal{B}(k)$ for every $A \subset [J]$, and by Theorem 3.4 it follows that

$$n_{NP}(k,\alpha) \sim 4\, N_A^*(k,\alpha) \quad \forall\, A \subset [J].$$

That is, for small values of α, the ESS increases by roughly a factor of 4 when utilizing the MNP rule, instead of the proposed asymptotically optimal Sum-Intersection rule. From Theorem 3.4 it also follows that for any $A \subset [J]$ we have

$$\liminf_{\alpha \to 0} \frac{n^*(k,\alpha)}{N_A^*(k,\alpha)} \geq \frac{4\mathcal{B}(k)}{\mathcal{B}(2k-1)}.$$

If in addition the hypotheses have identical information numbers, i.e., (3.11) holds, this lower bound is always larger than 2, which means that *any* fixed-sample size scheme will require at least twice as many observations as the Sum-Intersection rule, for small error probabilities.

3.4 Generalized familywise error rates *of both kinds*

In this section we study Problem 3.2. While we follow similar ideas and the results are of similar nature as in the previous section, the proposed procedure and the proof of its asymptotic optimality turn out to be much more complicated.

To describe the proposed multiple testing procedure, we first need to introduce some additional notations. Specifically, we denote by

$$0 < \widehat{\lambda}^1(n) \leq \ldots \leq \widehat{\lambda}^{p(n)}(n)$$

the order statistics of *positive* LLRs at time n, $\{\lambda^j(n) : \lambda^j(n) > 0,\, j \in [J]\}$, where $p(n)$ is the number of strictly positive LLRs at time n. Similarly, we denote by

$$0 \leq \breve{\lambda}^1(n) \leq \ldots \leq \breve{\lambda}^{q(n)}(n)$$

the order statistics of the absolute values of *non-positive* LLRs at time n, i.e., $\{-\lambda^j(n) : \lambda^j(n) \leq 0,\, j \in [J]\}$, where $q(n) := J - p(n)$. We also adopt the following convention:

$$\widehat{\lambda}^j(n) = \infty \;\text{ if }\; j > p(n), \quad \text{and} \quad \breve{\lambda}^j(n) = \infty \;\text{ if }\; j > q(n). \tag{3.22}$$

Moreover, we use the following notation

$$\lambda^{\widehat{i}_j(n)}(n) := \widehat{\lambda}^j(n), \quad \forall\, j \in \{1, \ldots, p(n)\},$$
$$\lambda^{\breve{i}_j(n)}(n) := -\breve{\lambda}^j(n), \quad \forall\, j \in \{1, \ldots, q(n)\},$$

for the indices of streams with *positive* and *non-positive* LLRs at time n, respectively. Thus, stream $\widehat{i}_1(n)$ (resp. $\breve{i}_1(n)$) has the least significant positive (resp. negative) LLR at time n.

3.4.1 Asymmetric Sum-Intersection rule

We begin by modifying the stopping rule, but not the decision rule, of the Sum-Intersection procedure (Subsection 3.3.1), in order to account for the asymmetry in the error metric that we consider in this section. This suggests a procedure $\delta_0(a,b) = (\tau_0, D_0)$ that stops as soon as the following two conditions are satisfied simultaneously: (i) the sum of the k_1 least significant positive LLRs is larger than $b > 0$, and (ii) the sum of the k_2 least significant negative LLRs is smaller than $-a < 0$. Formally,

$$\begin{aligned} \tau_0 &:= \inf\left\{ n \geq 1 : \sum_{j=1}^{k_1} \widehat{\lambda}^j(n) \geq b \ \text{ and } \ \sum_{j=1}^{k_2} \breve{\lambda}^j(n) \geq a \right\}, \\ D_0 &:= \left\{ j \in [J] : \ \lambda^j(\tau_0) > 0 \right\} = \left\{ \widehat{i}_1(\tau_0), \ldots, \widehat{i}_{p(\tau_0)}(\tau_0) \right\}, \end{aligned} \tag{3.23}$$

Similarly to the Sum-Intersection rule, this procedure, to which we refer as *asymmetric Sum-Intersection rule*, does not require strong evidence from every individual stream in order to terminate sampling. Indeed, upon stopping there may be insufficient evidence for the hypotheses that correspond to the $k_1 - 1$ least significant positive statistics and the $k_2 - 1$ least significant negative statistics, making them the anticipated false positives and false negatives, respectively, which we are allowed to make.

We will see that while the asymmetric Sum-Intersection rule can control generalized familywise error rates of both types, it will not in general be asymptotically optimal. To understand why this is the case, let A denote true subset of streams with signals and suppose that there is a subset B of ℓ streams with *noise*, i.e., $B \subset A^c$ with $|B| = \ell$, such that $\ell < k_1$ and

$$\mathcal{I}_1^j \ \gg \ \mathcal{I}_0^{i_1} \ \gg \ \mathcal{I}_0^{i_2}, \quad \forall\, j \in A, \quad i_1 \in A^c \setminus B, \quad i_2 \in B,$$

i.e., the hypotheses in streams with signal are much easier than in streams with noise, and the hypotheses in B are much harder than in the other streams with noise. In this case, the first stopping requirement in τ_0 will be easily satisfied, but not the second one, since the streams in B will slow down the growth of the sum of the k_2 least significant negative LLRs.

These observations suggest that, in the above scenario, the performance of δ_0 can be improved if we essentially "give up" the testing problems in B, in the sense that we presume that we make $\ell < k_1$ false positives for testing problems in B. This can be achieved by (i) ignoring the ℓ least significant negative

statistics in the second stopping requirement of τ_0, and asking the sum of the *next* k_2 least significant negative statistics to be small upon stopping, and (ii) modifying the decision rule to reject the nulls not only in streams with positive LLR, but also in the ℓ streams with the least significant *negative* LLRs upon stopping.

However, if we modify the decision rule in this way, we have spent from the beginning ℓ of the $k_1 - 1$ false positives we are allowed to make. This implies that we need to also modify the first stopping requirement in τ_0 and ask the sum of the $k_1 - \ell$ least significant positive LLRs to be large upon stopping.

If we denote by $\widehat{\delta}_\ell := (\widehat{\tau}_\ell, \widehat{D}_\ell)$ the procedure that incorporates the above modifications, then

$$\widehat{\tau}_\ell := \inf\left\{ n \geq 1 : \sum_{j=1}^{k_1-\ell} \widehat{\lambda}^j(n) \geq b \ \text{ and } \ \sum_{j=\ell+1}^{\ell+k_2} \check{\lambda}^j(n) \geq a \right\},$$
$$\widehat{D}_\ell := \{\widehat{i}_1(\widehat{\tau}_\ell), \ldots, \widehat{i}_{p(\widehat{\tau}_\ell)}(\widehat{\tau}_\ell)\} \bigcup \{\check{i}_1(\widehat{\tau}_\ell), \ldots, \check{i}_\ell(\widehat{\tau}_\ell)\},$$

where we omit the dependence on a, b in order to lighten the notation.

By the same token, if there are $\ell < k_2$ streams *with signal* in which the testing problems are much harder than in other streams, it is reasonable to expect that δ_0 may be outperformed by a procedure $\check{\delta}_\ell := (\check{\tau}_\ell, \check{D}_\ell)$, where

$$\check{\tau}_\ell := \inf\left\{ n \geq 1 : \sum_{i=\ell+1}^{\ell+k_1} \widehat{\lambda}^i(n) \geq b \ \text{ and } \ \sum_{j=1}^{k_2-\ell} \check{\lambda}^j(n) \geq a \right\}$$
$$\check{D}_\ell := \{\widehat{i}_{\ell+1}(\check{\tau}_\ell), \ \ldots \ , \widehat{i}_{p(\check{\tau}_\ell)}(\check{\tau}_\ell)\}.$$

Figure 3.1 provides a visualization of these stopping rules.

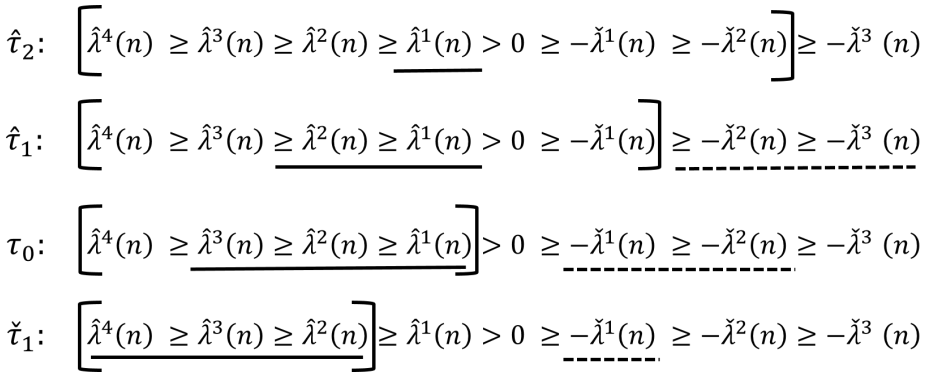


Figure 3.1: Set $J = 7$, $k_1 = 3$, $k_2 = 2$. Suppose at time n, $p(n) = 4, q(n) = 3$. Each rule stops when the sum of the terms with solid underline exceeds b, and at the same time the sum of the terms with dashed underline is below $-a$. Upon stopping, the null hypothesis for the streams in the bracket are rejected. Note that by convention (3.22), $\check{\lambda}^4(n) = \infty$, which makes the stopping rule $\hat{\tau}_2$ have only one condition to satisfy.

3.4.2 The Leap rule

The previous discussion suggests that the asymmetric Sum-Intersection rule, defined in (3.23), may be significantly outperformed by some of the procedures, $\{\widehat{\delta}_\ell, 0 \leq \ell < k_1\}$ and $\{\overset{\vee}{\delta}_\ell, 1 \leq \ell < k_2\}$, under some signal configurations, when the multiple testing problem is *asymmetric and/or inhomogeneous*. In this case, we propose combining the above procedures, i.e., stop as soon as any of them does so, and use the corresponding decision rule upon stopping. If multiple stopping criteria are satisfied at the same time, we then use the decision rule that rejects the most null hypotheses.

Formally, the proposed procedure $\delta_L := (T_L, D_L)$ is defined as follows:

$$\begin{aligned} T_L &:= \min\left\{\min_{0\leq\ell<k_1} \widehat{\tau}_\ell,\ \min_{1\leq\ell<k_2} \overset{\vee}{\tau}_\ell\right\}, \\ D_L &:= \left(\bigcup_{0\leq\ell<k_1, \widehat{\tau}_\ell = T_L} \widehat{D}_\ell\right) \bigcup \left(\bigcup_{1\leq\ell<k_2, \overset{\vee}{\tau}_\ell = T_L} \overset{\vee}{D}_\ell\right), \end{aligned} \tag{3.24}$$

and we refer to it as "Leap rule", because $\widehat{\delta}_\ell$ (resp. $\overset{\vee}{\delta}_\ell$) "leap" across the ℓ least significant negative (resp. positive) LLRs.

3.4.3 Error control of the Leap rule

We now show that the Leap rule can control generalized familywise error rates of both types.

Theorem 3.5. *Assume* (3.5) *holds. For any* $\alpha, \beta \in (0,1)$ *we have that* $\delta_L \in \Delta_{k_1,k_2}(\alpha,\beta)$ *when the thresholds are selected as follows:*

$$a = |\log(\beta)| + \log(2^{k_2} C_{k_2}^J), \quad b = |\log(\alpha)| + \log(2^{k_1} C_{k_1}^J). \tag{3.25}$$

Proof. The proof can be found in Section 3.10.1. □

The above threshold values are sufficient for establishing the asymptotic optimality of the Leap rule, but may be conservative in practice. Thus, as in the previous section, we recommend using simulation to find the thresholds that attain the target error probabilities. This means simulating for every $A \subset [J]$ the error probabilities of the Leap rule, $\mathsf{P}_A(|D_L(a,b) \setminus A| \geq k_1)$ and $\mathsf{P}_A(|A \setminus D_L(a,b)| \geq k_2)$, for various pairs of thresholds, a and b, and selecting the values for which the maxima (with respect to A) of the above error probabilities match the nominal levels, α and β, respectively.

As in the previous section, this task is facilitated when the multiple testing problem has a special structure. Specifically, when it is *symmetric* and the user-specified parameters are selected so that $\alpha = \beta$

and $k_1 = k_2$, i.e., when condition (3.12) holds, then we can select without any loss of generality the thresholds to be equal ($a = b$). If the multiple testing problem is *homogeneous*, the discussion following Theorem 3.1 also applies here.

3.4.4 Asymptotic optimality

For any $B \subset [J]$ and $1 \leq \ell \leq u \leq J$, we denote by

$$\mathcal{I}_1^{(1)}(B) \leq \ldots \leq \mathcal{I}_1^{(|B|)}(B)$$

the increasingly ordered sequence of $\mathcal{I}_1^j, j \in B$, and by

$$\mathcal{I}_0^{(1)}(B) \leq \ldots \leq \mathcal{I}_0^{(|B|)}(B)$$

the increasingly ordered sequence of $\mathcal{I}_0^j, j \in B$, and we set

$$\mathcal{D}_1(B;\ell,u) := \sum_{j=\ell}^{u} \mathcal{I}_1^{(j)}(B), \quad \text{where} \quad \mathcal{I}_1^{(j)}(B) = \infty \quad \text{for} \quad j > |B|,$$
$$\mathcal{D}_0(B;\ell,u) := \sum_{j=\ell}^{u} \mathcal{I}_0^{(j)}(B), \quad \text{where} \quad \mathcal{I}_0^{(j)}(B) = \infty \quad \text{for} \quad j > |B|.$$

The following lemma provides an asymptotic upper bound on the expected sample size of the stopping times that compose the stopping time of the Leap rule.

Lemma 3.1. *Assume* (3.8) *holds. For any* $A \subset [J]$ *we have as* $a, b \to \infty$

$$\mathsf{E}_A[\widehat{\tau}_\ell] \leq \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A;1,k_1-\ell)} \, , \, \frac{a(1+o(1))}{\mathcal{D}_0(A^c;\ell+1,\ell+k_2)} \right\}, \; 0 \leq \ell < k_1,$$
$$\mathsf{E}_A[\check{\tau}_\ell] \leq \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A;\ell+1,\ell+k_1)} \, , \, \frac{a(1+o(1))}{\mathcal{D}_0(A^c;1,k_2-\ell)} \right\}, \; 0 \leq \ell < k_2.$$

Proof. The proof can be found in Section 3.10.2. □

If thresholds are selected according to (3.25), then the upper bounds in the previous lemma take the following form

$$\widehat{L}_A(\ell;\alpha,\beta) := \max\left\{ \frac{|\log\alpha|}{\mathcal{D}_1(A;1,k_1-\ell)} \, , \, \frac{|\log\beta|}{\mathcal{D}_0(A^c;\ell+1,\ell+k_2)} \right\} \quad \text{for } \ell < k_1,$$
$$\check{L}_A(\ell;\alpha,\beta) := \max\left\{ \frac{|\log\alpha|}{\mathcal{D}_1(A;\ell+1,\ell+k_1)} \, , \, \frac{|\log\beta|}{\mathcal{D}_0(A^c;1,k_2-\ell)} \right\} \quad \text{for } \ell < k_2,$$

and from the definition of Leap rule in (3.24) it follows that as $\alpha, \beta \to 0$ we have $\mathsf{E}_A[T_L] \leq L_A(k_1, k_2, \alpha, \beta)\,(1+o(1))$, where

$$L_A(k_1, k_2, \alpha, \beta) := \min\left\{ \min_{0\leq \ell < k_1} \widehat{L}_A(\ell; \alpha, \beta)\ ,\ \min_{0\leq \ell < k_2} \breve{L}_A(\ell; \alpha, \beta) \right\}. \tag{3.26}$$

In the next theorem we show that it is not possible to achieve a smaller ESS, to a first-order asymptotic approximation as $\alpha, \beta \to 0$, proving in this way the asymptotic optimality of the Leap rule.

Theorem 3.6. *Assume* (3.8) *holds and that the thresholds in the Leap rule are selected such that* $\delta_L \in \Delta_{k_1,k_2}(\alpha, \beta)$ *and* $a \sim |\log(\beta)|, b \sim |\log(\alpha)|$, *e.g. according to* (3.25). *Then, for any* $A \subset [J]$ *we have as* $\alpha, \beta \to 0$,

$$\mathsf{E}_A\left[T_L\right] \ \sim\ L_A(k_1, k_2, \alpha, \beta) \ \sim\ N_A^*(k_1, k_2, \alpha, \beta).$$

Proof. In view of the discussion prior to the theorem, it suffices to show that for any $A \subset [J]$ we have as $\alpha, \beta \to 0$ that

$$N_A^*(k_1, k_2, \alpha, \beta) \geq L_A(k_1, k_2, \alpha, \beta)\,(1 - o(1)).$$

For the proof of this asymptotic lower bound we employ similar ideas as in the proof of Theorem 3.2 in the previous section. The change-of-measure argument is more complicated now, due to the interplay of the two kinds of error. We carry out the proof in Section 3.10.4. □

Remark 3.4. *When* $k_1 = k_2 = 1$, *the asymptotic optimality of the Intersection rule was established in [66] only in the i.i.d. case. Since the Leap rule coincides with the Intersection rule when* $k_1 = k_2 = 1$, *Theorem 3.6 generalizes this result in [66] beyond the i.i.d. case.*

We motivated the Leap rule by the inadequacy of the asymmetric Sum-Intersection rule, δ_0, in the case of *asymmetric and/or inhomogeneous* testing problems. In the following corollary we show that δ_0 is asymptotically optimal when (i) condition (3.11) holds, which is the case when the multiple testing problem is symmetric and homogeneous, and also (ii) the user-specified parameters are selected in a symmetric way, i.e., when (3.12) holds. In the same setup we establish the asymptotic optimality of the Intersection rule, δ_I, defined in (3.10).

Corollary 3.2. *Suppose* (3.8), (3.11), (3.12) *hold and consider the asymmetric Sum-Intersection rule* $\delta_0(b, b)$ *with* $b = b_\alpha$ *and the Intersection rule* $\delta_I(b, b)$ *with* $b = b_\alpha/k_1$, *where* b_a *is defined in* (3.19) *with* $k = k_1$. *Then*

$\delta_0, \delta_I \in \Delta_{k_1,k_1}(\alpha,\alpha)$, *and for any* $A \subset [J]$ *we have as* $\alpha \to 0$ *that*

$$\mathsf{E}_A[\tau_0] \sim \mathsf{E}_A[T_I] \sim \frac{|\log(\alpha)|}{k_1 \mathcal{I}} \sim N_A^*(k_1, k_1, \alpha, \alpha).$$

Proof. The proof can be found in Section 3.10.5. □

Remark 3.5. *In Section 3.5.2 we will illustrate numerically that when condition* (3.11) *is violated, both* δ_0 *and* δ_I *fail to be asymptotically optimal.*

3.4.5 Fixed-sample size rules

We now focus on the i.i.d. case (3.15) and consider procedures that stop at a *deterministic* time, which is selected to control the generalized familywise error rates.

For simplicity of presentation, we restrict ourselves to *homogeneous* testing problems, i.e., there are densities f_0 and f_1 such that

$$f_0^j = f_0, \quad f_1^j = f_1 \ \text{ for every } j \in [J]. \tag{3.27}$$

This assumption allows us to omit the dependence on the stream index j and write $\mathcal{I}_0 := \mathcal{I}_0^j$, $\mathcal{I}_1 := \mathcal{I}_1^j$ and $\Phi := \Phi^j$, where Φ^j is defined in (3.17). Moreover, we can apply the MNP rule, (3.14), without loss of generality, with the same threshold for each stream.

We further assume that user-specified parameters are selected as follows

$$k_1 = k_2, \quad \alpha = \beta^d \ \text{ for some } d > 0, \tag{3.28}$$

and that for each $d > 0$ there exists some $h_d \in (-\mathcal{I}_0, \mathcal{I}_1)$ such that

$$\Phi(h_d)/d = \Phi(h_d) - h_d. \tag{3.29}$$

When $d = 1$, condition (3.28) reduces to (3.12) and h_d is equal to 0. However, when $d \neq 1$, we allow for an asymmetric treatment of the two kinds of error.

Theorem 3.7. *Consider the multiple testing problem* (3.27) *and assume that the Kullback-Leibler numbers in* (3.16) *are positive and finite. Further, assume that* (3.28) *and* (3.29) *hold. Then as* $\beta \to 0$,

$$\frac{d\,(1-o(1))}{(2k_1-1)\Phi(h_d)} \leq \frac{n^*(k_1,k_1,\beta^d,\beta)}{|\log(\beta)|} \leq \frac{\widehat{n}_{NP}(k_1,k_1,\beta^d,\beta)}{|\log(\beta)|} \sim \frac{d}{k_1\Phi(h_d)}.$$

Proof. The proof is similar to that of Theorem 3.4, but it requires a *generalization* of Chernoff's lemma [20, Corollary 3.4.6] to account for the asymmetry of the two kinds of error. This generalization is presented in Lemma 3.15 and more details can be found in Section 3.10.6. □

Theorem 3.7, in conjunction with Theorem 3.6, allows us to quantify the performance loss that is induced by stopping at a deterministic time. To be more specific, we specialize the comparison in the case of testing the normal means (Example 3.1). By (3.18) we have $\mathcal{I} = \mathcal{I}_1 = \mathcal{I}_0$ and that for any $d \geq 1$

$$h_d = \frac{\sqrt{d}-1}{\sqrt{d}+1}\,\mathcal{I}, \quad \Phi(h_d) = \frac{d}{(1+\sqrt{d})^2}\,\mathcal{I},$$

and by Theorem 3.6 it follows that as $\beta \to 0$,

$$N_A^*(k_1, k_1, \beta^d, \beta) \sim L_A(k_1, k_1, \beta^d, \beta) \leq \widehat{L}_A(0; \beta^d, \beta) = \begin{cases} \frac{|\log(\beta)|}{k_1 \mathcal{I}}, & \text{if } |A| < k_1 \\ \\ \frac{d|\log(\beta)|}{k_1 \mathcal{I}}, & \text{if } |A| \geq k_1. \end{cases}$$

When in particular $d = 1$, for any $A \subset [J]$ we have

$$\begin{aligned} 2\, N_A^*(k_1, k_1, \beta, \beta)(1 - o(1)) &\leq n^*(k_1, k_1, \beta, \beta) \\ &\leq \widehat{n}_{NP}(k_1, k_1, \beta, \beta) \sim 4\, N_A^*(k_1, k_1, \beta, \beta), \end{aligned}$$

which agrees with the corresponding findings in Subsection 3.3.5.

3.5 Simulations for generalized familywise error rates

In this section we present two simulation studies that complement our asymptotic optimality theory in Section 3.4 for procedures that control generalized familywise error rates. The goal of the first study is to compare the proposed Leap rule (3.24) with the Intersection rule (3.10) and the asymmetric Sum-Intersection rule (3.23), in a *symmetric and homogeneous* setup where conditions (3.11) and (3.12) hold and all three procedures are asymptotically optimal. The goal of the second simulation study is to compare the same procedures when condition (3.11) is slightly violated, and only the Leap rule enjoys the asymptotic optimality property.

In both studies we consider the testing of normal means (Example 3.1), with $\sigma_j = 1$ for every $j \in [J]$. This is a *symmetric* multiple testing problem, where the Kullback-Leibler information in the j-th testing problem is $\mathcal{I}^j = \mu_j^2/2$. Moreover, we assume that condition (3.12) holds, i.e., $\alpha = \beta$ and $k_1 = k_2$. This

implies that we can set the thresholds in each *sequential* procedure to be equal, i.e., $a = b$, and as a result the two types of generalized familywise error rates will be the same. Finally, in both studies we include the performance of the fixed-sample size *multiple Neyman-Pearson* (MNP) rule (3.14), for which the choice of thresholds depends crucially on whether the problem is homogeneous or not.

In what follows, the error probability (Err) means the generalized familywise error rate of false positives (3.3), i.e., the *maximum* probability of k_1 false positives, with maximum being taken over all signal configurations. Thus, Err *does not* depend on the true subset of signals $A \subset [J]$.

3.5.1 Homogeneous case

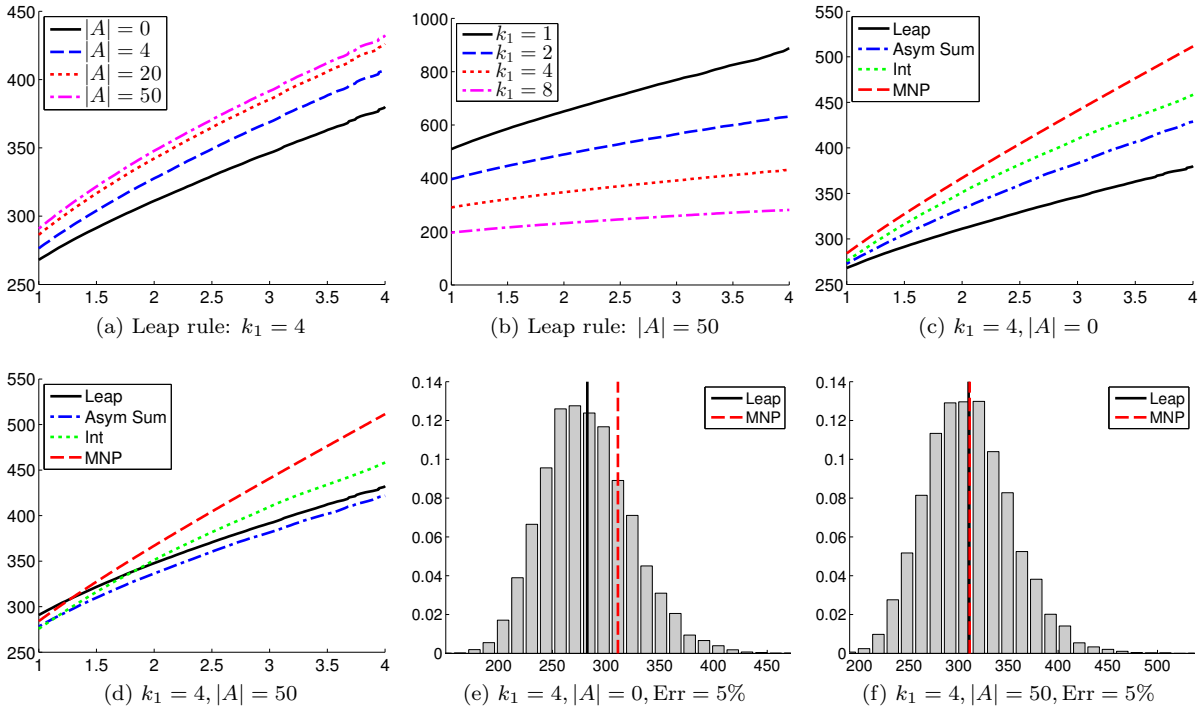
In the first simulation study we set $\mu_j = 0.25$ for each $j \in [J]$. In this homogeneous setup, the expected sample size (ESS) of all procedures under consideration depend only on the *number* of signals, and we can set the thresholds in the MNP rule, defined in (3.14), to be equal to 0. Moreover, it suffices to study the performance when the number of signals is no more than $J/2$. We consider $J = 100$ in Figure 3.2 and $J = 20$ in Figure 3.3.

In Figure 3.2a, we fix $k_1 = 4$ and evaluate the ESS of the Leap rule for four different cases regarding the number of signals. We see that, for any given Err, the smallest possible ESS is achieved in the boundary case of no signals ($|A| = 0$). This is because some components in the Leap rule only have one condition to be satisfied in the boundary cases (e.g. $\widehat{\tau}_2$ in Figure 3.1).

In Figure 3.2b, we fix the number of signals to be $|A| = 50$ and evaluate the Leap rule for different values of k_1. We observe that there are significant savings in the ESS as k_1 increases and more mistakes are tolerated.

In Figure 3.2c and 3.2d, we fix $k_1 = 4$ and compare the four rules for $|A| = 0$ and 50, respectively. In this *symmetric and homogeneous* setup, where (3.11) and (3.12) both hold, we have shown that all three sequential procedures are asymptotically optimal. Our simulations suggest that in practice the Leap rule works better when the number of signals, $|A|$, is close to 0 or J, but may perform slightly worse than the asymmetric Sum-Intersection rule, δ_0, when $|A|$ is close to $J/2$.

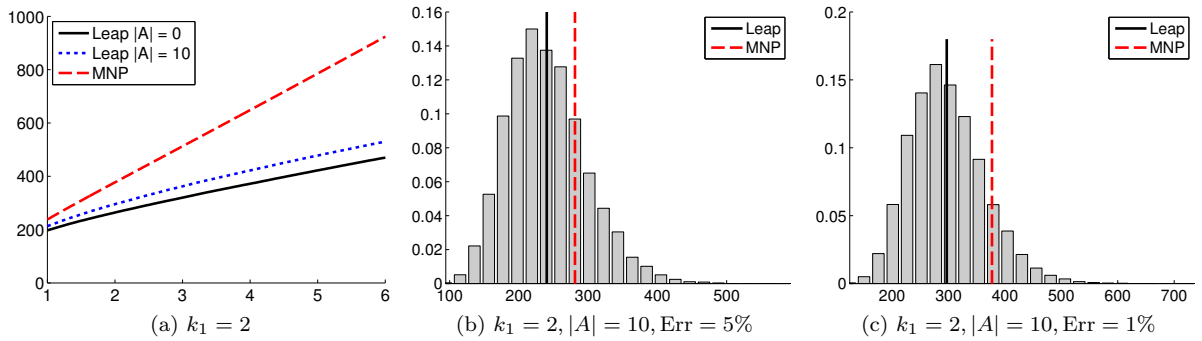
In Figure 3.2c, 3.2d and 3.3a, we also compare the performance of the Leap rule with the MNP rule. Further, in Figure 3.2e, 3.2f, 3.3b and 3.3c, we show the sampling distribution of the stopping time of the Leap rule at particular error levels. From these figures we can see that the best-case scenario for the MNP is when both the number of hypotheses, J, and the error probabilities, Err, are large. Note that this does not contradict our asymptotic analysis, where J is fixed and we let Err go to 0.



(a) Leap rule: $k_1 = 4$ (b) Leap rule: $|A| = 50$ (c) $k_1 = 4, |A| = 0$

(d) $k_1 = 4, |A| = 50$ (e) $k_1 = 4, |A| = 0, \text{Err} = 5\%$ (f) $k_1 = 4, |A| = 50, \text{Err} = 5\%$

Figure 3.2: Homogeneous case: $J = 100, k_1 = k_2$. In (a)-(d), the x-axis is $|\log_{10}(\text{Err})|$ and the y-axis is the ESS under P_A. In (e) and (f) are the sample distribution of the stopping time of the Leap rule with $\text{Err} = 5\%$.



(a) $k_1 = 2$ (b) $k_1 = 2, |A| = 10, \text{Err} = 5\%$ (c) $k_1 = 2, |A| = 10, \text{Err} = 1\%$

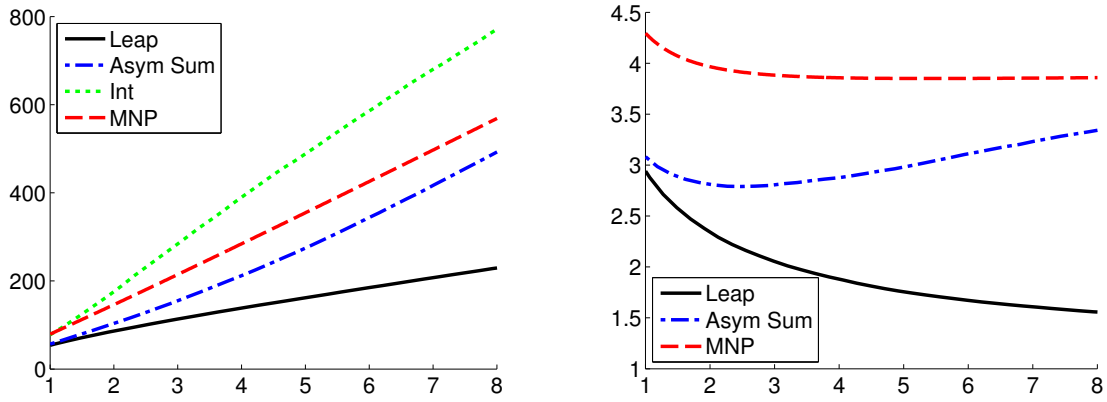
Figure 3.3: Homogeneous case: $J = 20, k_1 = 2$. In (a), the x-axis is $|\log_{10}(\text{Err})|$ and the y-axis is the ESS under P_A. In (b) and (c) are the sampling distribution of the stopping time of the Leap rule with $\text{Err} = 5\%$ and 1%.

3.5.2 Non-homogenous case

In the second simulation study we set $J = 10$, $\mu_j = 1/6$, $j = 1, 2$, $\mu_j = 1/2$, $j \geq 3$, so that the first two hypotheses are much harder than others. Specifically, $\mathcal{I}^j = 1/72$ for $j = 1, 2$, and $\mathcal{I}^j = 1/8$ for $j \geq 3$.

When the true subset of signals is $A^* = \{6, \cdots, 10\}$, the optimal asymptotic performance, (3.26), is equal to $8|\log(\text{Err})|$. In Figure 3.4a, we plot the ESS against $|\log_{10}(\text{Err})|$, and the ratio of ESS over $8|\log(\text{Err})|$ in Figure 3.4b. For the (asymptotically optimal) Leap rule, this ratio tends to 1 as $\alpha \to 0$. In contrast, the other rules have a different "slope" from the Leap rule in Figure 3.4a, which indicates that they fail to be asymptotically optimal in this context.

Finally, we note that in such a non-homogeneous setup, the choice of thresholds for the MNP rule (3.14) is not obvious. We found that instead of setting $h_j = 0$ for every $j \in [J]$, it is much more efficient to take advantage of the flexibility of generalized familywise error rates, as we did in the construction of the Leap rule in Subsection 3.4.2, and set $h_1 = -\infty$, $h_2 = \infty$ and $h_j = 0$ for $j \geq 3$. This choice "gives up" the first two "difficult" streams by always rejecting the null in the first one and accepting it in the second. The error constraints can then still be met as long as we do not make any mistakes in the remaining "easy" streams. In fact, we see that while the MNP rule behaves significantly worse than the asymptotically optimal Leap rule, it performs better than the Intersection rule, which "insists" on collecting strong enough evidence from each individual stream.



(a) ESS under P_{A^*} (b) Normalized by $8|\log(\text{Err})|$

Figure 3.4: Non-homogeneous case: $J = 10, k_1 = k_2 = 2, A^* = \{6, \cdots, 10\}$. The x-axis in both graphs is $|\log_{10}(\text{Err})|$. The y-axis in (a) is the ESS under P_{A^*}, and in (b) is the ratio of the ESS over $8|\log(\text{Err})|$.

3.6 Extension to composite hypotheses

We now extend the setup introduced in Section 3.2, allowing both the null and the alternative hypothesis in each local testing problem to be composite. Thus, for each $j \in [J]$, the distribution of X^j, the sequence of observations in the j-th stream, is now parametrized by $\theta^j \in \Theta^j$, where Θ^j is a subset of some Euclidean space, and the hypothesis testing problem in the j-th stream becomes

$$\mathsf{H}_0^j : \theta^j \in \Theta_0^j \quad \text{versus} \quad \mathsf{H}_1^j : \theta^j \in \Theta_1^j,$$

where Θ_0^j and Θ_1^j are two disjoint subsets of Θ^j. When $A \subset [J]$ is the subset of streams in which the alternative is correct, we denote by $\boldsymbol{\Theta}_A$ the subset of the parameter space $\boldsymbol{\Theta} := \Theta^1 \times \ldots \times \Theta^J$ that is compatible with A, i.e.,

$$\boldsymbol{\Theta}_A := \{(\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta} : \ \theta^j \in \Theta_1^j \ \Leftrightarrow \ j \in A\}.$$

We denote by $\mathsf{P}_{\theta^j}^j$ the distribution of the j-th stream when the value of its local parameter is θ^j. Moreover, we denote by $\mathsf{P}_{A,\boldsymbol{\theta}}$ the underlying probability measure when the subset of signals is A and the parameter is $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta}_A$, and by $\mathsf{E}_{A,\boldsymbol{\theta}}$ the corresponding expectation. Due to the independence across streams, we have $\mathsf{P}_{A,\boldsymbol{\theta}} = \mathsf{P}_{\theta^1}^1 \otimes \ldots \otimes \mathsf{P}_{\theta^J}^J$.

Our presentation in the case of composite hypotheses will focus on the control of generalized *familywise error rates*; the corresponding treatment of the generalized mis-classification rate will be similar. Thus, given $k_1, k_2 \geq 1$ and $\alpha, \beta \in (0,1)$, the class of procedures of interest now is:

$$\Delta_{k_1,k_2}^{comp}(\alpha,\beta) := \{(T,D) : \max_{A,\,\boldsymbol{\theta}} \mathsf{P}_{A,\boldsymbol{\theta}}(|D \setminus A| \geq k_1) \leq \alpha \quad \text{and} \\ \max_{A,\,\boldsymbol{\theta}} \mathsf{P}_{A,\boldsymbol{\theta}}(|A \setminus D| \geq k_2) \leq \beta\},$$

and the goal is the same as the one in Problem 3.2 with $N_A^*(k_1,k_2,\alpha,\beta)$ being replaced by

$$N_{A,\boldsymbol{\theta}}^*(k_1,k_2,\alpha,\beta) := \inf_{(T,D)\in\Delta_{k_1,k_2}^{comp}(\alpha,\beta)} \mathsf{E}_{A,\boldsymbol{\theta}}[T],$$

and the asymptotic optimality being achieved for every $A \subset [J]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}_A$.

3.6.1 Leap rule with adaptive log-likelihood ratios

The proposed procedure in this setup is a modification of the Leap rule (3.24), where the local LLR statistics are replaced by statistics that account for the composite nature of the two hypotheses. To be more specific,

for every $j \in [J]$ and $n \in \mathbb{N}$ we denote by $\ell^j(n, \theta^j)$ the log-likelihood function (with respect to some σ-finite measure ν_n^j) in the j-th stream based on the first n observations, i.e.,

$$\ell^j(n, \theta^j) := \ell^j(n-1, \theta^j) + \log\left(p^j_{\theta^j}(X^j(n)\,|\mathcal{F}^j_{n-1})\right); \quad \ell^j(0, \theta^j) := 0,$$

where $p^j_{\theta^j}(X^j(n)\,|\mathcal{F}^j_{n-1})$ is the conditional density of $X^j(n)$ given the previous $n-1$ observations in the j-th stream. Moreover, for every stream $j \in [J]$ and time $n \in \mathbb{N}$ we denote by $\ell^j_i(n)$ the corresponding *generalized* log-likelihood under $\mathsf{H}^{\mathsf{j}}_{\mathsf{i}}$, i.e.,

$$\ell^j_i(n) := \sup\left\{\ell^j(n, \theta^j) \;:\; \theta^j \in \Theta^j_i\right\}, \quad i = 0, 1.$$

Further, at each $n \in \mathbb{N}$, we select an $\mathcal{F}_n$-measurable estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}^1_n, \ldots, \widehat{\theta}^J_n) \in \boldsymbol{\Theta}$, and define the *adaptive log-likelihood* statistic for the j-th stream as follows:

$$\ell^j_*(n) := \ell^j_*(n-1) + \log\left(p^j_{\widehat{\theta}^j_{n-1}}(X^j(n)\,|\mathcal{F}^j_{n-1})\right); \quad \ell^j_*(0) = 0, \tag{3.30}$$

where $\widehat{\boldsymbol{\theta}}_0 := (\widehat{\theta}^1_0, \ldots, \widehat{\theta}^J_0) \in \boldsymbol{\Theta}$ is some deterministic initialization. The proposed procedure in this context is the Leap rule (3.24), where each LLR statistic $\lambda^j(n)$ is replaced by the following *adaptive* log-likelihood ratio:

$$\lambda^j_*(n) := \begin{cases} \ell^j_*(n) - \ell^j_0(n), & \text{if } \ell^j_0(n) < \ell^j_1(n) \text{ and } \ell^j_0(n) < \ell^j_*(n) \\ -(\ell^j_*(n) - \ell^j_1(n)), & \text{if } \ell^j_1(n) < \ell^j_0(n) \text{ and } \ell^j_1(n) < \ell^j_*(n) \\ \text{undefined}, & \text{otherwise}, \end{cases} \tag{3.31}$$

with the understanding that there is no stopping at time n if $\lambda^j_*(n)$ is undefined for some j. Clearly, large positive values of λ^j_* support $\mathsf{H}^{\mathsf{j}}_{\mathsf{1}}$, whereas large negative values of λ^j_* support $\mathsf{H}^{\mathsf{j}}_{\mathsf{0}}$. We denote this modified version of the Leap rule by $\delta^*_L(a, b) = (T^*_L, D^*_L)$.

In the next subsection we establish the asymptotic optimality of δ^*_L under general conditions. In Section 3.11.5 we discuss in more detail the above adaptive statistics, as well as other choices for the local statistics. In Section 3.11.4 we demonstrate with a simulation study that if we replace the LLR λ^j by the adaptive statistic λ^j_* (3.31) in the *Intersection rule* (3.10) and the *asymmetric Sum-Intersection rule* (3.23), then these procedures fail to be asymptotically optimal *even in the presence of special structures*. Finally, we should point out that the gains over fixed-sample size procedures will also be larger compared to the case

of simple hypotheses, as sequential methods are, by definition, adaptive to the true parameter.

3.6.2 Asymptotic optimality

First of all, for each $j \in [J]$ we generalize condition (3.7) and assume that for any distinct $\theta^j, \tilde{\theta}^j \in \Theta^j$ there exists a positive number $I^j(\theta^j, \tilde{\theta}^j)$ such that

$$\frac{1}{n}\left(\ell^j(n,\theta^j) - \ell^j(n,\tilde{\theta}^j)\right) \xrightarrow[n\to\infty]{\mathsf{P}^j_{\theta^j} \text{ completely}} I^j(\theta^j,\tilde{\theta}^j). \tag{3.32}$$

Second, we require that the null and alternative hypotheses in each stream are separated, in the sense that if for each $j \in [J]$ and $\theta^j \in \Theta^j$ we define

$$\mathcal{I}_0^j(\theta^j) := \inf_{\tilde{\theta}^j \in \Theta_1^j} I^j(\theta^j,\tilde{\theta}^j) \qquad \text{and} \qquad \mathcal{I}_1^j(\theta^j) := \inf_{\tilde{\theta}^j \in \Theta_0^j} I^j(\theta^j,\tilde{\theta}^j), \tag{3.33}$$

then we assume that

$$\mathcal{I}_0^j(\theta^j) > 0 \ \ \forall\, \theta^j \in \Theta_0^j \qquad \text{and} \qquad \mathcal{I}_1^j(\theta^j) > 0 \ \ \forall\, \theta^j \in \Theta_1^j. \tag{3.34}$$

Finally, we assume that for each $j \in [J]$ and $\epsilon > 0$,

$$\begin{aligned} &\sum_{n=1}^{\infty} \mathsf{P}^j_{\theta^j}\left(\frac{\ell^j_*(n) - \ell^j_1(n)}{n} - \mathcal{I}_0^j(\theta^j) < -\epsilon\right) < \infty \text{ for every } \theta^j \in \Theta_0^j, \\ &\sum_{n=1}^{\infty} \mathsf{P}^j_{\theta^j}\left(\frac{\ell^j_*(n) - \ell^j_0(n)}{n} - \mathcal{I}_1^j(\theta^j) < -\epsilon\right) < \infty \text{ for every } \theta^j \in \Theta_1^j. \end{aligned} \tag{3.35}$$

We now state the main result of this section, the asymptotic optimality of δ^*_L under the above conditions. The proof is presented in Section 3.11.

Theorem 3.8. *Assume* (3.32), (3.34) *and* (3.35) *hold. Further, assume the thresholds in the Leap rule are selected such that* $\delta^*_L(a,b) \in \Delta^{comp}_{k_1,k_2}(\alpha,\beta)$ *and* $a \sim |\log(\beta)|, b \sim |\log(\alpha)|$, *e.g. according to* (3.25). *Then, for any* $A \subset [J]$ *and* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_A$, *we have as* $\alpha, \beta \to 0$,

$$\mathsf{E}_{A,\boldsymbol{\theta}}\left[T_L\right] \ \sim \ L_{A,\boldsymbol{\theta}}(k_1,k_2,\alpha,\beta) \ \sim \ N^*_{A,\boldsymbol{\theta}}(k_1,k_2,\alpha,\beta),$$

where $L_{A,\boldsymbol{\theta}}(k_1,k_2,\alpha,\beta)$ *is a quantity defined in Section 3.11.1 that characterizes the asymptotic optimal performance.*

While conditions (3.32) and (3.34) are easily satisfied and simple to check, the one-sided complete convergence condition (3.35) is not as apparent. It is known [71, p. 278-280] that when $\widehat{\theta}_n^j$ is selected to be the Maximum Likelihood estimator (MLE) of θ^j, condition (3.35) is satisfied when testing a normal mean with unknown variance, as well as when testing the coefficient of a first-order autoregressive model. In Section 3.12 we further show that condition (3.35) is satisfied when (i) the data in each stream are i.i.d. with some *multi-parameter exponential family* distribution, and (ii) the null and the alternative parameter spaces are compact.

3.7 Conclusion

In this Chapter we have considered the sequential multiple testing problem under two error metrics. In the first one, the goal is to control the probability of at least k mistakes, of any kind. In the second one, the goal is to control simultaneously the probabilities of at least k_1 false positives and at least k_2 false negatives. Assuming that the data for the various hypotheses are obtained sequentially in independent streams, we characterized the optimal performance to a first-order asymptotic approximation as the error probabilities vanish, and proposed the first asymptotically optimal procedure for each of the two problems. Procedures that are asymptotically optimal under classical error control ($k = 1$, $k_1 = k_2 = 1$) were found to be suboptimal under *generalized* error metrics apart from very special cases. Moreover, in the case of i.i.d. data streams, we quantified the asymptotic savings in the expected sample size relative to fixed-sample size procedures.

There are certain questions that remain open. First, we conducted a first-order asymptotic analysis, ignoring higher-order terms in the approximation to the optimal performance. The latter however appears to be non-negligible in practice(see Figure 3.4b). Thus, it is an open problem to obtain a more precise characterization of the optimal performance, as well as to examine whether the proposed rules enjoy a stronger optimality property. Second, the number of streams is treated as constant in our asymptotic analysis, but can be very large in practice. It is interesting to consider an enhanced asymptotic regime, where the number of streams also goes to infinity as the error probabilities vanish. Third, although simulation techniques can be used to determine threshold values that guarantee the error control, it is desirable to have closed-form expressions for less conservative threshold values.

There are several generalizations. One direction is to relax the assumption that the streams corresponding to the different testing problems are independent. Another direction is to allow for early stopping in some streams, in which case the goal is to minimize the total number of observations in all streams. Finally, it is

interesting to study FDR-type error control.

3.8 Simulations for generalized mis-classification rate

In this section, we present two simulation studies that complement our asymptotic optimality theory for procedures that control the generalized mis-classification rate (Section 3.3). Specifically, our goal is to compare the proposed Sum-Intersection rule and the Intersection rule in two setups. The first one is a *symmetric and homogeneous* setup, in which (3.11) holds and both rules are asymptotically optimal. The second one is a non-homogeneous setup, where the condition (3.11) is (slightly) violated and the Intersection rule fails to be asymptotically optimal. In each setup, we also include the performance of the multiple Neyman-Pearson rule (MNP) (3.14), which is a fixed-sample size procedure.

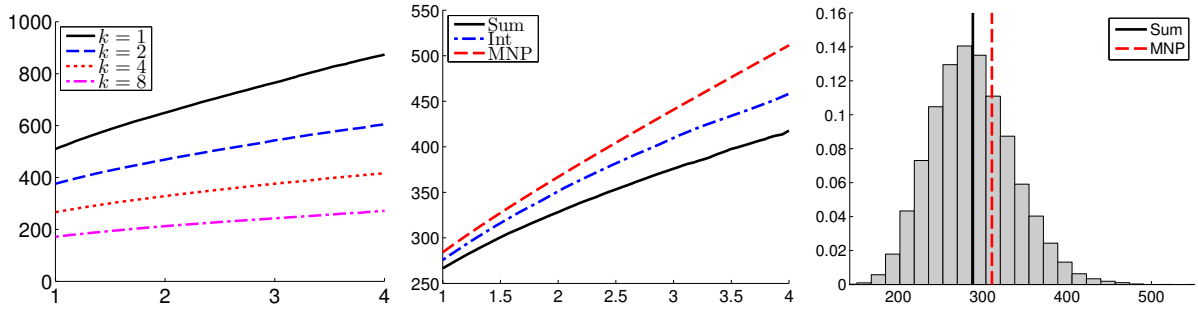
For these comparisons, we consider the testing of normal means, introduced in Example 3.1. As discussed in Example 3.1, this problem is *symmetric*. As a result, we set $h = 0$ in the MNP rule (3.14), and further the performance of each rule under consideration is the same for any subset of signals. Thus we do not need to specify the actual subset of signals.

3.8.1 Homogeneous case

We set in Example 3.1 $\mu_j = 0.25, \sigma_j = 1$ for $j \in [J]$. We consider $J = 100$ in Figure 3.5 and $J = 20$ in Figure 3.6.

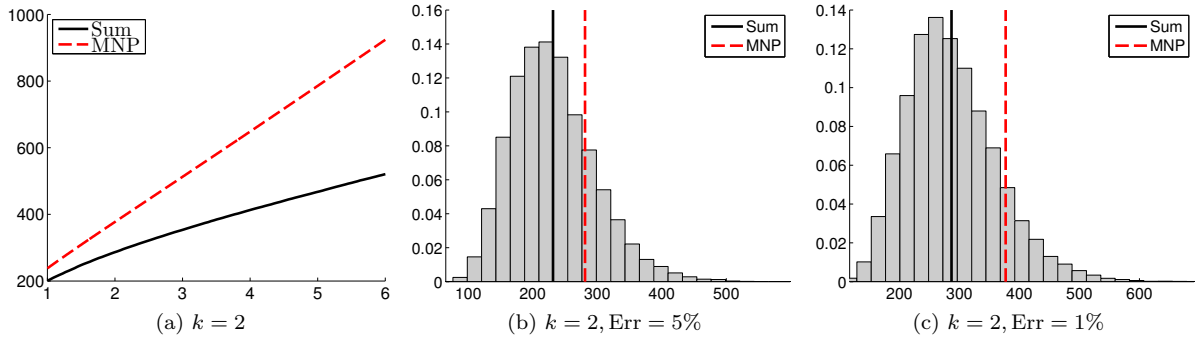
In Figure 3.5a, we study the performance of the Sum-Intersection rule for different values of k. We observe that there are significant savings in the ESS as k increases and more mistakes are tolerated. In Figure 3.5b, we compare the three rules for $k = 4$. Although both sequential rules enjoy the asymptotic optimality property in this setup, we observe that the Sum-Intersection rule outperforms the Intersection rule in terms of ESS.

In Figure 3.5b and 3.6a, we also compare the Sum-Intersection rule with the MNP rule. Further, in Figure 3.5c, 3.6b and 3.6c, we show the sampling distribution of the Sum-Intersection at particular error levels. From these figures, we observe that the advantage of sequential procedures over the MNP rule is significant if J is not too large or Err is small.



(a) Sum rule: vary k (b) $k = 4$, vary rules (c) $k = 4, \text{Err} = 5\%$

Figure 3.5: Homogeneous case: $J = 100$. In (a) and (b), the x-axis is $|\log_{10}(\text{Err})|$ and the y-axis represents the ESS. In (c), we study the sample distribution of the stopping time of the Sum-Intersection rule with $\text{Err} = 5\%$.



(a) $k = 2$ (b) $k = 2, \text{Err} = 5\%$ (c) $k = 2, \text{Err} = 1\%$

Figure 3.6: Homogeneous case: $J = 20$. In (a), the x-axis is $|\log_{10}(\text{Err})|$ and the y-axis represents the ESS. In (b) and (c), we study the sample distribution of the stopping time of the Sum-Intersection rule with $\text{Err} = 5\%$ and 1%.

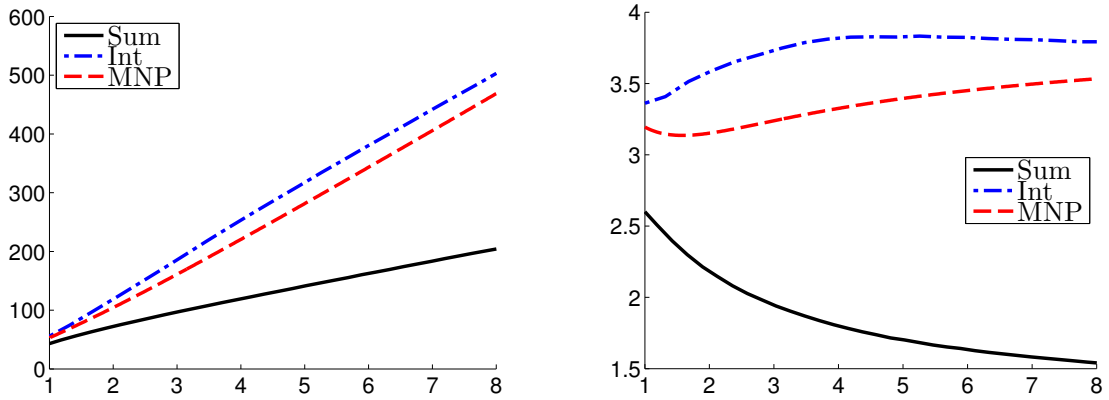
3.8.2 Non-homogeneous case

Second, we set $J = 10$, $k = 2$ and

$$f_0^j = \mathcal{N}(0,1) \;\; \forall\, j \in [J], \quad f_1^j = \begin{cases} \mathcal{N}(1/6,1) & \text{if } \; j = 1 \\ \mathcal{N}(1/2,1) & \text{if } \; j \geq 2 \end{cases} .$$

In this second setup, we have injected a slight violation of homogeneity. All testing problems are identical apart from the first one, which is much harder than the other ones. Indeed, $\mathcal{I}_0^j = \mathcal{I}_1^j = \mathcal{I}^j$, where $\mathcal{I}^j = 1/72$ for $j = 1$, and $\mathcal{I}^j = 1/8$ for $j \geq 2$. Since $k = 2$, the optimal asymptotic performance in this problem is determined by the two most difficult hypotheses and is equal to $7.2|\log(\text{Err})|$. In Figure 3.7a we plot the expected sample size(ESS) against $|\log_{10}(\text{Err})|$ and in Figure 3.7b we plot the ratio of ESS over $7.2|\log(\text{Err})|$.

We observe that this ratio tends to 1 for the asymptotically optimal Sum-Intersection rule, whereas this is not the case for the other two rules. In particular, as predicted by Theorem 3.4, the ratio for the MNP rule tends to 4.



(a) ESS vs $|\log_{10}(\text{Err})|$ (b) Normalized version

Figure 3.7: Non-homogeneous case: $J = 10, k = 2$. The x-axis in both graphs is $|\log_{10}(\text{Err})|$. The y-axis is the corresponding ESS in (a), and is the ratio of the ESS over $7.2|\log(\text{Err})|$ in (b).

3.9 Proofs regarding the generalized mis-classification rate

3.9.1 Proofs of Theorem 3.1

Proof. It suffices to show that for any $b > 0$ and $A \subset [J]$ we have

$$\mathsf{P}_A(|A \triangle D_S(b)| \geq k) \leq C_k^J \, e^{-b}.$$

Fix $A \subset [J]$ and $b > 0$. Observe that the event $\{|A \triangle D_S| \geq k\}$ occurs if and only if there exist $B_1 \subset A$ and $B_2 \subset A^c$ such that $|B_1| + |B_2| = k$ and the following event occurs:

$$\Gamma(B_1, B_2) := \left\{ D_S^i = 0,\, D_S^j = 1,\, \forall\, i \in B_1, j \in B_2 \right\}.$$

Since there are C_k^J such pairs, due to Boole's inequality it suffices to show that the probability of each of these events is bounded by e^{-b}. To this end, fix $B_1 \subset A, B_2 \subset A^c$ such that $|B_1| + |B_2| = k$ and consider the set $C = (A \setminus B_1) \cup B_2$. Then, with the change of measure $\mathsf{P}_A \to \mathsf{P}_C$, we have

$$\mathsf{P}_A(\Gamma(B_1, B_2)) = \mathsf{E}_C \left[\exp\left\{ \lambda^{A,C}(T_S) \right\}; \Gamma(B_1, B_2) \right]. \tag{3.36}$$

For $i \in B_1$ we have $D_S^i = 0$, which implies $\lambda^i(T_S) \leq 0$, and for $j \in B_2$ we have $D_S^j = 1$, which implies $\lambda^j(T_S) > 0$. Thus, on the event $\Gamma(B_1, B_2)$,

$$\begin{aligned} \lambda^{A,C}(T_S) &= \sum_{i \in B_1} \lambda^i(T_S) - \sum_{j \in B_2} \lambda^j(T_S) \\ &= -\sum_{i \in B_1} |\lambda^i(T_S)| - \sum_{j \in B_2} |\lambda^j(T_S)| \leq -\sum_{i=1}^{k} \widetilde{\lambda}^i(T_S) \leq -b, \end{aligned} \tag{3.37}$$

where the first equality is due to (2.3), the first inequality follows from the definition of $\widetilde{\lambda}^i$'s, and the second from the definition of the stopping time T_S. Thus, the proof is complete in view of (3.36). □

3.9.2 An important Lemma

The following lemma is crucial in establish Theorem 3.2.

Lemma 3.2. *Let $A, B \subset [J]$. Then there exists $B^* \subset [J]$ such that*

$$(i) \quad B \notin \mathcal{U}_k(B^*), \qquad (ii) \quad I^{A,B^*} \leq \mathcal{D}_A(k).$$

To show Lemma 3.2, we start with a lemma about sets.

Lemma 3.3. *Let $A, B, \Gamma \subset [J]$. There exists $B^* \subset [J]$ such that*

$$A \triangle B^* \quad \subset \quad \Gamma \quad \subset \; B \triangle B^*$$

Proof. Define the following disjoint sets:

$$B_1 = B \cap \Gamma, \quad B_2 = B^c \cap \Gamma, \quad A_1 = A \cap \Gamma^c, \quad A_2 = A^c \cap \Gamma^c$$

Clearly, $\Gamma = B_1 \cup B_2$, and $\Gamma^c = A_1 \cup A_2$. Let $B^* = B_2 \cup A_1$.

On one hand, if $j \in B_1$, then $j \in B$ and $j \notin B^*$; if $j \in B_2$, then $j \notin B$ and $j \in B^*$. It implies $\Gamma = B_1 \cup B_2 \subset B \triangle B^*$.

On the other, if $j \in A_1$, then $j \in A$ and $j \in B^*$; if $j \in A_2$, then $j \notin A$ and $j \notin B^*$. Thus $\Gamma^c = A_1 \cup A_2 \subset (A \triangle B^*)^c$, which implies $A \triangle B^* \subset \Gamma$. □

Now we are ready to prove Lemma 3.2.

Proof. Let $C^* \notin \mathcal{U}_k(A)$ such that $\mathcal{D}_A(k) = \mathcal{I}^{A,C^*}$ and set $\Gamma = A \triangle C^*$. Then, clearly $|\Gamma| \geq k$. By Lemma 3.3, there exists a set $B^* \subset [J]$ such that

$$A \triangle B^* \quad \subset \quad \Gamma = A \triangle C^* \quad \subset \quad B \triangle B^*.$$

From the second inclusion it follows that $|B \triangle B^*| \geq |\Gamma| \geq k$, which proves (i). From the first inclusion it follows that $A \setminus B^* \subset A \setminus C^*$ and $B^* \setminus A \subset C^* \setminus A$, therefore from (2.3) we conclude that

$$\mathcal{I}^{A,B^*} = \sum_{i \in A \setminus B^*} \mathcal{I}_1^i + \sum_{j \in B^* \setminus A} \mathcal{I}_0^j \leq \sum_{i \in A \setminus C^*} \mathcal{I}_1^i + \sum_{j \in C^* \setminus A} \mathcal{I}_0^j = \mathcal{I}^{A,C^*},$$

which proves (ii). □

3.9.3 Proof of Theorem 3.2

Proof. Fix $A \subset [J]$, $k \in [J]$, and set

$$\ell_\alpha := |\log(\alpha)| / \mathcal{D}_A(k), \quad \alpha \in (0,1).$$

By Markov's inequality, for any stopping time T, $\alpha \in (0,1)$ and $q > 0$,

$$\mathsf{E}_A[T] \geq q\ell_\alpha \, \mathsf{P}_A(T \geq q\ell_\alpha).$$

Thus, it suffices to show for every $q \in (0,1)$ we have

$$\liminf_{\alpha \to 0} \inf_{(T,D) \in \Delta_k(\alpha)} \mathsf{P}_A(T \geq q\ell_\alpha) \geq 1, \tag{3.38}$$

as this will imply $\liminf_{\alpha \to 0} N_A^*(k,\alpha)/\ell_\alpha \geq q$, and the desired result will follow by letting $q \to 1$.

In order to prove (3.38), let us start by fixing arbitrary $\alpha, q \in (0,1)$ and $(T,D) \in \Delta_k(\alpha)$. Then,

$$1 - \alpha \leq \mathsf{P}_A(D \in \mathcal{U}_k(A)) = \sum_{B \in \mathcal{U}_k(A)} \mathsf{P}_A(D = B). \tag{3.39}$$

Now, consider an arbitrary $B \in \mathcal{U}_k(A)$, and let $B^* \subset [J]$ be a set that satisfies the two conditions in Lemma 3.2. Then, $|B^* \triangle B| \geq k$, and consequently

$$\mathsf{P}_{B^*}(D = B) \leq \alpha. \tag{3.40}$$

We can now decompose the probability $\mathsf{P}_A(D = B)$ as follows:

$$\mathsf{P}_A\left(\lambda^{A,B^*}(T) < \log\left(\frac{\eta}{\alpha}\right); D = B\right) + \mathsf{P}_A\left(\lambda^{A,B^*}(T) \geq \log\left(\frac{\eta}{\alpha}\right); D = B\right),$$

where η is an arbitrary constant in $(0,1)$. We denote the first term by I and second by II. For the first term, by a change of measure $\mathsf{P}_A \to \mathsf{P}_{B^*}$ we have

$$\begin{aligned} \mathrm{I} &= \mathsf{E}_{B^*}\left[\exp\{\lambda^{A,B^*}(T)\}\ ;\ \lambda^{A,B^*}(T) < \log\left(\frac{\eta}{\alpha}\right), D = B\right] \\ &\leq \frac{\eta}{\alpha}\mathsf{P}_{B^*}(D = B) \leq \eta, \end{aligned}$$

where the second inequality follows from (3.40). For the second term, we have

$$\mathrm{II} \leq \mathsf{P}_A\left(T \leq q\frac{|\log\alpha|}{\mathcal{D}_A(k)},\ \lambda^{A,B^*}(T) \geq \log\left(\frac{\eta}{\alpha}\right)\right) + \mathsf{P}_A(T \geq q\ell_\alpha, D = B).$$

By construction, B^* satisfies $\mathcal{I}^{A,B^*} \leq \mathcal{D}_A(k)$; thus the first term in the right-hand side is bounded above by

$$\epsilon_{\alpha,B^*} := \mathsf{P}_A\left(T \leq q\frac{|\log\alpha|}{I^{A,B^*}},\quad \lambda^{A,B^*}(T) \geq |\log\alpha| + \log(\eta)\right).$$

Due to the SLLN (3.6), we have

$$\mathsf{P}_A\left(\lim_{n\to\infty}\frac{\lambda^{A,B^*}(n)}{n}=\mathcal{I}^{A,B^*}\right)=1.$$

Therefore, by Lemma 3.13, it follows that $\epsilon_{\alpha,B^*}\to 0$ as $\alpha\to 0$.

Putting everything together we have

$$\mathsf{P}_A(D=B)\le\eta+\epsilon_{\alpha,B^*}+\mathsf{P}_A(T\ge q\ell_\alpha, D=B),$$

and summing over $B\in\mathcal{U}_k(A)$ we obtain

$$\begin{aligned}\mathsf{P}_A(D\in\mathcal{U}_k(A))&\le|\mathcal{U}_k(A)|\eta+\epsilon_\alpha+\mathsf{P}_A(T\ge q\ell_\alpha, D\in\mathcal{U}_k(A))\\&\le|\mathcal{U}_k(A)|\eta+\epsilon_\alpha+\mathsf{P}_A(T\ge q\ell_\alpha),\end{aligned}$$

where $\epsilon_\alpha:=\sum_{B\in\mathcal{U}_k(A)}\epsilon_{\alpha,B^*}\to 0$ as $\alpha\to 0$. Due to (3.39), we have

$$\mathsf{P}_A(T\ge q\ell_\alpha)\ge\ 1-\alpha-\epsilon_\alpha-|\mathcal{U}_k(A)|\eta.$$

Since $(T,D)\in\Delta_k(\alpha)$ is arbitrary and $\alpha\in(0,1)$ also arbitrary, taking the infimum over (T,D) and letting $\alpha\to 0$ we obtain

$$\liminf_{\alpha\to 0}\ \inf_{(T,D)\in\Delta_k(\alpha)}\mathsf{P}_A(T\ge q\ell_\alpha)\ge 1-|\mathcal{U}_k(A)|\eta.$$

Finally, letting $\eta\to 0$ we obtain (3.38), which completes the proof. □

3.9.4 Proof of Theorem 3.3

The following fact about set operations will be needed:

$$\text{Let }\ A,B\subset[J]\ \text{ and }\ C=A\triangle B\ .\ \text{Then } A\triangle C=B. \tag{3.41}$$

Proof. Fix $A\subset[J]$ and consider the stopping time

$$T^A(b):=\inf\left\{n\ge 1:\lambda^{A,C}(n)\ge b\quad\forall\, C\notin\mathcal{U}_k(A)\right\}.$$

Under the conditions of the lemma, from Lemma 3.14 in the Section it follows that $b \to \infty$ we have

$$\mathsf{E}_A[T^A(b)] \leq \frac{b\,(1+o(1))}{\mathcal{D}_A(k)}.$$

Thus, it suffices to show that $T_S(b) \leq T^A(b)$ for any given $b > 0$. In what follows, we fix $b > 0$ and suppress the dependence on b. By the definition of the Sum-Intersection rule, it suffices to show that

$$\sum_{i \in B} |\lambda^i(T^A)| \geq b, \quad \forall\, B \subset [J] :\ |B| = k. \tag{3.42}$$

To this end, fix $B \subset [J]$ with $|B| = k$ and set $C = A \triangle B$. Then, from (3.41) we have that $B = A \triangle C$. Since $|B| \geq k$, it follows that $C \notin \mathcal{U}_k(A)$, and by the definition of T^A we have $\lambda^{A,C}(T^A) \geq b$. As a result,

$$\begin{aligned} b \leq \lambda^{A,C}(T^A) &= \sum_{i \in A \setminus C} \lambda^i(T^A) - \sum_{j \in C \setminus A} \lambda^j(T^A) \\ &\leq \sum_{i \in A \,\triangle\, C} |\lambda^i(T^A)| = \sum_{i \in B} |\lambda^i(T^A)|. \end{aligned}$$

The proof is complete in view of (3.42). □

3.9.5 Proof of Corollary 3.1

Proof. Fix $A \subset [J]$. For (i) it suffices to show that for any $b > 0$

$$\mathsf{P}_A(|A \triangle D_I(b,b)|) \leq C_k^J\, e^{-kb}.$$

The proof is identical to that of Theorem 3.1 as long as we replace the inequalities in (3.37) by

$$-\sum_{i \in B_1} |\lambda^i(T_I)| - \sum_{j \in B_2} |\lambda^j(T_I)| \leq -kb.$$

In order to prove (ii), setting $k = 1$ in Theorem 3.3 we have as $b \to \infty$

$$\mathsf{E}_A[T_I(b,b)] \leq \frac{b\,(1+o(1))}{\min_{C \neq A} \mathcal{I}^{A,C}}. \tag{3.43}$$

If condition (3.11) is satisfied, then $\min_{C \neq A} \mathcal{I}^{A,C} = \mathcal{I}$. Therefore, if $b \sim |\log \alpha|/k$, from (3.43) we have that as $\alpha \to 0$

$$\mathsf{E}_A\,[T_I] \leq \frac{|\log \alpha|}{k\mathcal{I}}(1+o(1)).$$

Further, this asymptotic upper bound agrees with the asymptotic lower bound in (3.20), since $\mathcal{D}_A(k) = k\mathcal{I}$ when condition (3.11) holds. Thus, the proof is complete. □

3.9.6 Proof of Theorem 3.4

Proof. Since $k \leq (J+1)/2$ is fixed, we write $n^*(\alpha)$ (resp. $n_{NP}(\alpha)$) for $n^*(k,\alpha)$ (resp. $n_{NP}(k,\alpha)$) for simplicity. By Theorem 3.3, for any $A \subset [J]$,

$$N_A^*(k,\alpha) \sim \frac{|\log \alpha|}{\mathcal{D}_A(k)}, \text{ as } \alpha \to 0. \tag{3.44}$$

(i) Let us first focus on $n^*(\alpha)$. By its definition (3.13), there exist some

$$D^*(\alpha) \in \Delta_{fix}(n^*(\alpha)) \cap \Delta_k(\alpha).$$

Denote P the probability measure for data in all streams. For any $A \subset [J]$ with $|A| = 2k-1$, we consider the following simple versus simple problem:

$$\mathsf{H}_0' : \mathsf{P} = \mathsf{P}_\emptyset \quad \text{vs.} \quad \mathsf{H}_1' : \mathsf{P} = \mathsf{P}_A, \tag{3.45}$$

where P_A is defined in (3.2). Consider the following procedure for (3.45):

$$\bar{D}^*(\alpha) = \begin{cases} 0 & \text{if } |D^*(\alpha)| < k \\ 1 & \text{if } |D^*(\alpha)| \geq k \end{cases}.$$

Then by definition of $D^*(\alpha)$, we have

$$\mathsf{P}_\emptyset(\bar{D}^*(\alpha) = 1) = \mathsf{P}_\emptyset(|D^*(\alpha)| \geq k) \leq \alpha,$$
$$\mathsf{P}_A(\bar{D}^*(\alpha) = 0) = \mathsf{P}_A(|D^*(\alpha)| < k) \leq \alpha,$$

where the second inequality uses the fact that $|A| = 2k-1$. Thus

$$\frac{1}{n^*(\alpha)} \log(\alpha) \geq \frac{1}{n^*(\alpha)} \log\left(\frac{1}{2}\mathsf{P}_\emptyset(\bar{D}^*(\alpha) = 1) + \frac{1}{2}\mathsf{P}_A(\bar{D}^*(\alpha) = 0)\right).$$

By Chernoff's lemma 3.15,

$$\liminf_{\alpha\to 0}\frac{1}{n^*(\alpha)}\log\left(\frac{1}{2}\mathsf{P}_\emptyset(\bar{D}^*(\alpha)=1)+\frac{1}{2}\mathsf{P}_A(\bar{D}^*(\alpha)=0)\right)\geq -\Phi^A(0)$$

where $\Phi^A(0):=\sup_{\theta\in\mathbb{R}}\left\{-\log\left(\mathsf{E}_\emptyset\left[e^{\theta\lambda^{A,\emptyset}(1)}\right]\right)\right\}$. Due to independence,

$$\Phi^A(0)=\sup_{\theta\in\mathbb{R}}\left\{\sum_{j\in A}-\log\left(\mathsf{E}_0^j\left[e^{\theta\lambda^j(1)}\right]\right)\right\}\leq\sum_{j\in A}\Phi^j(0).$$

As a result, we have

$$\liminf_{\alpha\to 0}\frac{1}{n^*(\alpha)}\log(\alpha)\geq-\sum_{j\in A}\Phi^j(0)=-\sum_{j\in A}\mathcal{C}_j,$$

By maximizing the lower bound over $A\subset[J]$ with $|A|=2k-1$, we have

$$\liminf_{\alpha\to 0}\frac{n^*(\alpha)}{|\log(\alpha)|}\geq\frac{1}{\sum_{j=1}^{2k-1}\mathcal{C}^{(j)}},$$

which, together with (3.44), completes the proof of (i).

(ii) Now let us focus on $n_{NP}(\alpha)$. By definition, there exists some $\tilde{h}\in\mathbb{R}^J$ such that

$$(n_{NP}(\alpha),\tilde{D}(\alpha))\in\Delta_k(\alpha),\ \text{where}\ \tilde{D}(\alpha):=D_{NP}(n_{NP}(\alpha),\tilde{h}).$$

Denote

$$\begin{aligned}p_j&:=\mathsf{P}_0^j(\tilde{D}^j(\alpha)=1)=\mathsf{P}_0^j\left(\frac{1}{n_{NP}(\alpha)}\lambda^j(n_{NP}(\alpha))>\tilde{h}_j\right)\\q_j&:=\mathsf{P}_1^j(\tilde{D}^j(\alpha)=0)=\mathsf{P}_1^j\left(\frac{1}{n_{NP}(\alpha)}\lambda^j(n_{NP}(\alpha))\leq\tilde{h}_j\right)\end{aligned}$$

For any $A_1,A_2\subset[J]$ such that $A_1\cap A_2=\emptyset$ and $|A_1\cup A_2|=k$,

$$\begin{aligned}\alpha&\geq\mathsf{P}_{A_1}\left(\cap_{j\in A_1}\{\tilde{D}^j(\alpha)=0\}\bigcap\cap_{i\in A_2}\{\tilde{D}^i(\alpha)=1\}\right)=\prod_{j\in A_1}q_j\prod_{i\in A_2}p_i,\\\alpha&\geq\mathsf{P}_{A_2}\left(\cap_{j\in A_1}\{\tilde{D}^j(\alpha)=1\}\bigcap\cap_{i\in A_2}\{\tilde{D}^i(\alpha)=0\}\right)=\prod_{j\in A_1}p_j\prod_{i\in A_2}q_i.\end{aligned}$$

Since A_1, A_2 are arbitrary, we have for any $A \subset [J]$ with $|A| = k$

$$\alpha \geq \prod_{j \in A} \max\{p_j, q_j\},$$

which implies that

$$\log(\alpha) \geq \sum_{j \in A} \max\{\log(p_j), \log(q_j)\} \geq \sum_{j \in A} \log(\frac{1}{2}p_j + \frac{1}{2}q_j).$$

Thus again by Chernoff's Lemma 3.15,

$$\liminf_{\alpha \to 0} \frac{1}{n_{NP}(\alpha)} \log(\alpha) \geq -\sum_{j \in A} \Phi^j(0).$$

Maximizing the lower bound over $A \subset [J]$ with $|A| = k$, we have

$$\liminf_{\alpha \to 0} \frac{n_{NP}(\alpha)}{|\log(\alpha)|} \geq \frac{1}{\sum_{j=1}^k \mathcal{C}^{(j)}}.$$

By the same argument, if we choose $\tilde{h} = 0$, the equality is achieved. Then the proof of (ii) is complete in view of (3.44). □

3.9.7 Bernoulli example under the generalized mis-classification rate

For simplicity, let us assume that for each $j \in [J]$, $\{X^j(n) : n \in \mathbb{N}\}$ are i.i.d. Bernoulli random variables, and the hypotheses are *homogeneous*. Thus, we assume that there exists some constant $p \in (0, 1/2)$ such that for each $j \in [J]$,

$$\mathsf{H}_0^j : \mathsf{P}_0^j(X^j(1) = 1) = p \text{ versus } \mathsf{H}_1^j : \mathsf{P}_1^j(X^j(1) = 1) = 1 - p := q.$$

In this case, $\mathcal{I} = \mathcal{I}_0^j = \mathcal{I}_1^j = H(p)$, where $H(x) := x \log(\frac{x}{1-x}) + (1-x)\log(\frac{1-x}{x})$. Further,

$$\Phi(0) = \sup_{\theta \in \mathbb{R}} \left\{ -\log(p^\theta q^{1-\theta} + p^{1-\theta} q^\theta) \right\} = \log \frac{1}{2\sqrt{p(1-p)}}.$$

By Theorem 3.4, for any $A \subset [J]$,

$$\liminf_{\alpha \to 0} \frac{n^*(k, \alpha)}{N_A^*(k, \alpha)} \geq \frac{kH(p)}{(2k-1)\Phi(0)}, \quad \lim_{\alpha \to 0} \frac{n_{NP}(k, \alpha)}{N_A^*(k, \alpha)} = \frac{H(p)}{\Phi(0)}.$$

In Figure 3.8, we plot $H(p)/\Phi(0)$ as a function of p.

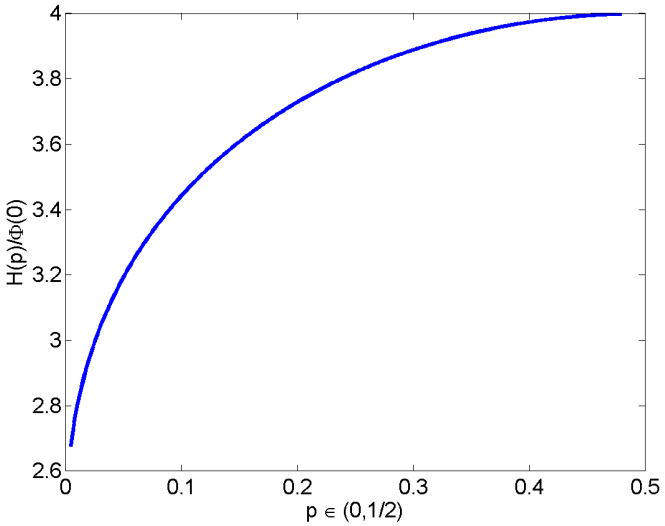


Figure 3.8: The plot for $H(p)/\Phi(0)$ as a function of p

3.10 Proofs regarding the generalized familywise error rates

3.10.1 Proof of Theorem 3.5

The goal in this subsection is to show that for any $a, b > 0$ and $A \subset [J]$ we have

$$\mathsf{P}_A(|D_L \setminus A| \geq k_1) \leq Q(k_1)\, e^{-b}, \quad \mathsf{P}_A(|A \setminus D_L| \geq k_2) \leq Q(k_2)\, e^{-a},$$

where $Q(k) = 2^k C_k^J$. We start with a lemma that shows how to select the thresholds for procedures $\widehat{\delta}_\ell$, $0 \leq \ell < k_1$ and $\breve{\delta}_\ell$, $0 \leq \ell < k_2$.

Lemma 3.4. *Assume that* (3.5) *holds. Fix* $A \subset [J]$. *Let* $B_1 \subset A^c$ *with* $|B_1| = k_1$, *and* $B_2 \subset A$ *with* $|B_2| = k_2$.

(i) Fix any $0 \leq \ell < k_1$. *For any event* $\Gamma \in \mathcal{F}_{\widehat{\tau}_\ell}$, *we have*

$$\mathsf{P}_A(B_1 \subset \widehat{D}_\ell) \leq C_\ell^{k_1} e^{-b}, \quad \mathsf{P}_A(B_2 \subset \widehat{D}_\ell^c, \Gamma) \leq e^{-a} \mathsf{P}_{A \setminus B_2}(\Gamma).$$

(ii) Fix any $0 \leq \ell < k_2$. *For any event* $\Gamma \in \mathcal{F}_{\breve{\tau}_\ell}$, *we have*

$$\mathsf{P}_A(B_1 \subset \breve{D}_\ell, \Gamma) \leq e^{-b} \mathsf{P}_{A \cup B_1}(\Gamma), \quad \mathsf{P}_A(B_2 \subset \breve{D}_\ell^c) \leq C_\ell^{k_2} e^{-a}.$$

Proof. We will only prove (i), since (ii) can be shown in a similar way. Fix $0 \leq \ell < k_1$. By definition, $\widehat{D}_\ell$ rejects the nulls in the ℓ streams with the least significant non-positive LLR, in addition to the nulls in the streams with positive LLR. Thus,

$$\{B_1 \subset \widehat{D}_\ell\} \subset \bigcup_{M \subset B_1, |M| = k_1 - \ell} \Pi_M \ , \quad \text{where } \Pi_M := \{\lambda^j(\widehat{\tau}_\ell) > 0 \ \ \forall j \ \in M\}.$$

With a change of measure from $\mathsf{P}_A \to \mathsf{P}_C$, where $C = A \cup M$, we have

$$\mathsf{P}_A(\Pi_M) = \mathsf{E}_C \left[\exp\{\lambda^{A,C}(\widehat{\tau}_\ell)\}; \Pi_M\right] = \mathsf{E}_C \left[\exp\left\{-\sum_{j \in M} \lambda^j(\widehat{\tau}_\ell)\right\}; \Pi_M\right].$$

By the definition of $\widehat{\tau}_\ell$, on the event Π_M we have $\sum_{j \in M} \lambda^j(\widehat{\tau}_\ell) \geq b$. Thus $\mathsf{P}_A(\Pi_M) \leq e^{-b}$. Since the number of such M is no more than $C_\ell^{k_1}$, the first inequality in (i) follows from Boole's inequality.

On the other hand, we observe that on the event $\{B_2 \subset \widehat{D}_\ell^c\}$, we have

$$\sum_{j \in B_2} \lambda^j(\widehat{\tau}_\ell) \leq -a.$$

Thus with a change of measure from $\mathsf{P}_A \to \mathsf{P}_{A \setminus B_2}$, we have

$$\mathsf{P}_A(B_2 \subset \widehat{D}_\ell^c, \Gamma) \leq \mathsf{E}_{A \setminus B_2} \left[\exp\left\{\sum_{j \in B_2} \lambda^j(\widehat{\tau}_\ell)\right\}; \Gamma\right] \leq e^{-a} \mathsf{P}_{A \setminus B_2}(\Gamma),$$

which completes the proof. □

Proof of Theorem 3.5. We will only establish the upper bound for $\mathsf{P}_A(|A \setminus D_L| \geq k_2)$, since the other inequality can be established similarly. Observe that

$$\{|A \setminus D_L| \geq k_2\} \ \subset \bigcup_{B \subset A : |B| = k_2} \{B \subset D_L^c\}.$$

Since the union consists at most $C_{k_2}^J$ events, by Boole's inequality, it suffices to show that the probability of each event is upper bounded by $2^{k_2} e^{-a}$. Fix an *arbitrary* $B \subset A$ with $|B| = k_2$. Further observe that

$$\{B \subset D_L^c\} \subset \cup_{\ell=0}^{k_1 - 1} \hat{\Gamma}_{B,\ell} \ \bigcup \ \cup_{\ell=1}^{k_2 - 1} \check{\Gamma}_{B,\ell}, \ \text{ where}$$
$$\widehat{\Gamma}_{B,\ell} := \{B \subset \widehat{D}_\ell^c\} \cap \{D_L = \widehat{D}_\ell\}, \quad \check{\Gamma}_{B,\ell} := \{B \subset \check{D}_\ell^c\}.$$

By Boole's inequality it follows that $\mathsf{P}_A(B \subset D_L^c)$ is upper bounded by

$$\sum_{\ell=0}^{k_1-1} \mathsf{P}_A(\hat{\Gamma}_{B,\ell}) + \sum_{\ell=1}^{k_2-1} \mathsf{P}_A(\check{\Gamma}_{B,\ell}) \leq \sum_{\ell=0}^{k_1-1} e^{-a} \mathsf{P}_{A\setminus B}(D_L = \hat{D}_\ell) + \sum_{\ell=1}^{k_2-1} C_\ell^{k_2} e^{-a}$$
$$\leq e^{-a} + e^{-a}\left(\sum_{\ell=1}^{k_2-1} C_\ell^{k_2}\right) \leq 2^{k_2} e^{-a},$$

where the first inequality follows from Lemma 3.4, and the second from the fact that $\{D_L = \hat{D}_\ell\}$ are disjoint events. Thus, the proof is complete. □

3.10.2 Proof of Lemma 3.1

Proof. We will only prove the inequality for $\widehat{\tau}_\ell$, as the proof of the inequality for $\check{\tau}_\ell$ is similar. Fix A and $0 \leq \ell < k_1$. We introduce the following classes of subsets

$$\mathcal{M}_1 = \{B \subset A : |B| = k_1 - \ell\},$$
$$\mathcal{M}_0 = \left\{B \subset A^c : |B| = k_2,\, \mathcal{I}_0^i \geq \mathcal{I}_0^{(\ell+1)}(A^c)\ \forall\, i \in B\right\}.$$

Clearly, we have $\widehat{\tau}_\ell \leq \tau'$, where

$$\tau' := \inf\{n \geq 1 : \min_{B\in\mathcal{M}_1} \sum_{i\in B} \lambda^i(n) \geq b \ \text{ and } \ \min_{B\in\mathcal{M}_0} \sum_{j\in B} \lambda^j(n) \leq -a,$$
$$\min_{i\in A} \lambda^i(n) > 0 \ \text{ and } \ \max_{j\notin A} \lambda^j(n) < 0\}.$$

Thus, by an application of Lemma 3.14, we have

$$\mathsf{E}_A[\tau'] \leq \max\left\{\frac{b}{\min_{B\in\mathcal{M}_1}\sum_{j\in B} I_1^j}\ ,\ \frac{a}{\min_{B\in\mathcal{M}_0}\sum_{j\in B} I_0^j}\right\}(1+o(1)).$$

By definition, for any $B_1 \in \mathcal{M}_1$ and $B_0 \in \mathcal{M}_0$, we have

$$\sum_{j\in B} I_1^j \geq \mathcal{D}_1(A; 1, k_1-\ell), \quad \sum_{j\in B} I_0^j \geq \mathcal{D}_0(A^c; 1+\ell, k_2+\ell)$$

therefore we conclude that

$$\mathsf{E}_A[\tau'] \leq \max\left\{\frac{b}{\mathcal{D}_1(A; 1, k_1-\ell)}\ ,\ \frac{a}{\mathcal{D}_0(A^c; 1+\ell, k_2+\ell)}\right\}(1+o(1)),$$

which proves the inequality for $\widehat{\tau}_\ell$. □

3.10.3 An important lemma

In this subsection, we establish a lemma that is critical in establishing the lower bound in Theorem 3.6. To state the result, let us denote by

$$\mathcal{U}_{k_1,k_2}(A) = \{C \subset [J] : |C \setminus A| < k_1 \text{ and } |A \setminus C| < k_2\}, \tag{3.46}$$

the collection of sets that are "close" to A, according to the generalized familywise error rates. Since k_1, k_2 are fixed integers, for simplicity of notations, we write in this subsection

$$L(A;\alpha,\beta) \quad \text{for} \quad L_A(k_1,k_2,\alpha,\beta).$$

Lemma 3.5. *Let $A \subset [J]$, $B \in \mathcal{U}_{k_1,k_2}(A)$, and $\alpha, \beta > 0$.*

1. *If $|B| \geq k_1$ and $|B^c| \geq k_2$, then there exists $B_1^*, B_2^* \subset [J]$ such that*

$$(i)\ |B \setminus B_1^*| = k_1,\ |B_2^* \setminus B| = k_2, \quad (ii)\ \frac{|\log(\alpha)|}{\mathcal{I}^{A,B_1^*}} \bigvee \frac{|\log(\beta)|}{\mathcal{I}^{A,B_2^*}} \geq L(A;\alpha,\beta)$$

2. *If $|B| < k_1$, then there exists $B_2^* \subset [J]$ such that*

$$(i)\ |B_2^* \setminus B| = k_2, \quad (ii)\ \frac{|\log(\beta)|}{\mathcal{I}^{A,B_2^*}} \geq L(A;\alpha,\beta).$$

3. *If $|B^c| < k_2$, there exists $B_1^* \subset [J]$ such that*

$$(i)\ |B \setminus B_1^*| = k_1, \quad (ii)\ \frac{|\log(\alpha)|}{\mathcal{I}^{A,B_1^*}} \geq L(A;\alpha,\beta).$$

The proof relies on the following two lemmas.

Lemma 3.6. *Let $G \subset A \subset F \subset [J]$. Denote $s_1 = |A \setminus G|$ and $s_2 = |F^c|$. Then for any integer n, we have*

$$\begin{aligned} D_1(G,1,n) &\leq D_1(A,1+s_1,n+s_1), \\ D_0(F \setminus A,1,n) &\leq D_0(A^c,1+s_2,n+s_2) \end{aligned}$$

Proof. Let's start with the first inequality. We can assume $n \leq |G|$, since otherwise both sides are equal to ∞.

Fix some $1 \leq i \leq n$. Then clearly the i^{th} smallest element in $\{\mathcal{I}_1^j : j \in G\}$ is no larger than the $(i + |A \setminus G|)^{th}$ element in $\{\mathcal{I}_1^j : j \in A\}$. Thus the first inequality follows from the definition of the D_1 function.

For the second inequality, it follows from the previous argument by replacing G by $F \setminus A$, A by A^c, and $\mathcal{I}_1^j$ by $\mathcal{I}_0^j$. □

Lemma 3.7. *Let ℓ_1, ℓ_2 be two non-negative integers such that $\ell_1 < k_1$ and $\ell_2 < k_2$. Then for any $A \subset [K]$, and $\alpha, \beta > 0$, we have*

$$\frac{|\log(\alpha)|}{\mathcal{D}_1(A, 1 + \ell_2, k_1 - \ell_1 + \ell_2)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, 1 + \ell_1, k_2 - \ell_2 + \ell_1)} \geq L(A; \alpha, \beta).$$

Proof. Let's consider the case that $\ell_1 \geq \ell_2$. When $\ell_1 \leq \ell_2$, the result can be proved in a similar way. Thus, denote $\ell = \ell_1 - \ell_2$. Then

$$\begin{aligned}
& \frac{|\log(\alpha)|}{\mathcal{D}_1(A, 1 + \ell_2, k_1 - \ell_1 + \ell_2)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, 1 + \ell_1, k_2 - \ell_2 + \ell_1)} \\
= & \frac{|\log(\alpha)|}{\mathcal{D}_1(A, 1 + \ell_2, k_1 - l)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, 1 + \ell + \ell_2, k_2 + \ell)} \\
\geq & \frac{|\log(\alpha)|}{\mathcal{D}_1(A, 1, k_1 - \ell)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, 1 + \ell, k_2 + \ell)} \\
= & \widehat{L}_A(\ell; \alpha, \beta) \quad \geq \quad L(A; \alpha, \beta)
\end{aligned}$$

where the last line used the definition of $\widehat{L}_A$ and L. □

With above two lemmas, we're ready to present the proof of Lemma 3.5. We illustrate the intuition of the following proof in Figure 3.9.

Proof. Fix A and $B \in \mathcal{U}_{k_1,k_2}(A)$. By definition of the class $\mathcal{U}_{k_1,k_2}(A)$,

$$\ell_1 := |B \setminus A| < k_1, \quad \ell_2 := |A \setminus B| < k_2.$$

First, consider the case that $|B| \geq k_1$, which implies $|A \cap B| \geq k_1 - \ell_1$. Thus we can find $\Gamma_1 \subset A \cap B$

such that

$$\Gamma_1 = k_1 - \ell_1, \quad \sum_{i\in\Gamma_1} \mathcal{I}_1^i = \mathcal{D}_1(A\cap B, 1, k_1-\ell_1)$$

Let's consider $B_1^* := A \setminus \Gamma_1$; it's easy to see

$$A \setminus B_1^* = \Gamma_1, \quad B \setminus B_1^* = \Gamma_1 \cup (B \setminus A)$$

Thus, $|B \setminus B_1^*| = k_1$; further, viewing $A \cap B$ as G in the Lemma 3.6, and since $\ell_2 = |A \setminus B|$, we have

$$\mathcal{I}^{A,B_1^*} = \sum_{i\in\Gamma_1} \mathcal{I}_1^i = \mathcal{D}_1(A\cap B, 1, k_1-\ell_1) \le \mathcal{D}_1(A, 1+\ell_2, k_1-\ell_1+\ell_2).$$

Second, consider the case that $|B^c| \ge k_2$, which implies $|A^c \cap B^c| \ge k_2 - \ell_2$. Thus there exists $\Gamma_2 \subset A^c \cap B^c$ such that

$$\Gamma_2 = k_2 - \ell_2, \quad \sum_{j\in\Gamma_2} I_0^j = \mathcal{D}_0(A^c\cap B^c, 1, k_2-\ell_2)$$

Let's consider $B_2^* := A \cup \Gamma_2$; it's easy to see

$$B_2^* \setminus A = \Gamma_2, \quad B_2^* \setminus B = \Gamma_2 \cup (A \setminus B)$$

Then $|B_2^* \setminus B| = k_2$. further, viewing $A \cup (A^c \cap B^c)$ as F in the Lemma 3.6, and since $\ell_1 = |B \setminus A| = |F^c|$, we have

$$\mathcal{I}^{A,B_2^*} = \sum_{j\in\Gamma_2} \mathcal{I}_0^j = \mathcal{D}_0(A^c\cap B^c, 1, k_2-\ell_2) \le \mathcal{D}_0(A^c, 1+\ell_1, k_2-\ell_2+\ell_1)$$

It remains to show B_1^* and B_2^* satisfy the property (ii) in each case.

Case 1: $|B| \ge k_1$ and $|B^c| \ge k_2$. By construction of B_1^* and B_2^*, we have

$$\begin{aligned} & \frac{|\log(\alpha)|}{\mathcal{I}^{A,B_1^*}} \bigvee \frac{|\log(\beta)|}{I^{A,B_2^*}} \\ \ge\ & \frac{|\log(\alpha)|}{\mathcal{D}_1(A, \ell_2+1, \ell_2+k_1-\ell_1)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, \ell_1+1, \ell_1+k_2-\ell_2)} \\ \ge\ & L(A;\alpha,\beta) \end{aligned}$$

where the last inequality is due to Lemma 3.7.

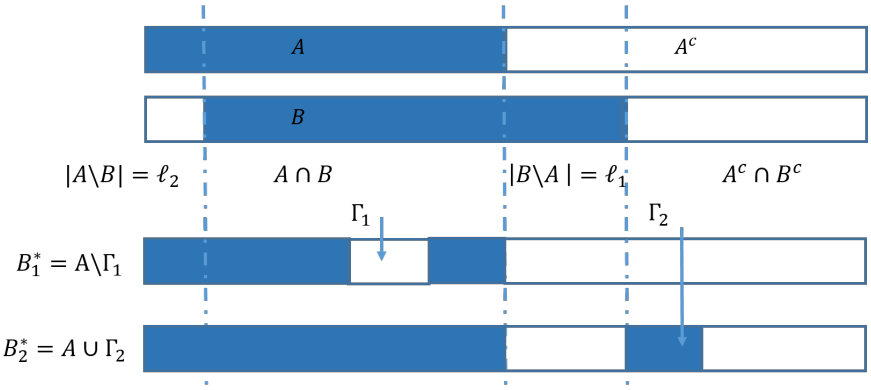


Figure 3.9: The solid area are the streams with signal. The whole set $[J]$ is partitioned into four disjoint sets: $A \setminus B$, $A \cap B$, $B \setminus A$, $A^c \cap B^c$. If $B \in \mathcal{U}_{k_1,k_2}(A)$, then $\ell_1 < k_1$ and $\ell_2 < k_2$.

Case 2: $|B| < k_1$, which implies the following:

$$|A| = |A \setminus B| + |A \cap B| = \ell_2 + |B| - \ell_1 < \ell_2 + k_1 - \ell_1$$

and thus $D_1(A, \ell_2 + 1, \ell_2 + k_1 - \ell_1) = \infty$. As a result,

$$\begin{aligned}\frac{|\log(\beta)|}{I^{A,B_2^*}} &\geq \frac{|\log(\alpha)|}{\mathcal{D}_1(A, \ell_2 + 1, \ell_2 + k_1 - \ell_1)} \bigvee \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, \ell_1 + 1, \ell_1 + k_2 - \ell_2)} \\ &\geq L(A; \alpha, \beta)\end{aligned}$$

where the last inequality is again due to Lemma 3.7.

Case 3: $|B^c| < k_2$. It can be proved in the same way as in case 2. □

3.10.4 Proof of Theorem 3.6

As explained in the discussion following Theorem 3.6, it suffices to show that for any $A \subset [J]$, as $\alpha, \beta \to 0$,

$$N_A^*(k_1, k_2, \alpha, \beta) \geq L_A(k_1, k_2, \alpha, \beta)\,(1 - o(1)).$$

Since k_1, k_2 are fixed integers, for simplicity of notations, we write in this subsection

$$L(A; \alpha, \beta) \quad \text{for} \quad L_A(k_1, k_2, \alpha, \beta).$$

Proof. Fix $A \subset [J]$. By the same argument as in the proof of Theorem 3.2, it suffices to show for every

$q \in (0,1)$ we have:

$$\liminf_{\alpha,\beta\to 0} \inf_{(T,D)\in\Delta_{k_1,k_2}(\alpha,\beta)} \mathsf{P}_A\left(T \geq qL(A;\alpha,\beta)\right) \geq 1.$$

Fix $q \in (0,1)$ and let (T,D) be any procedure in $\Delta_{k_1,k_2}(\alpha,\beta)$. Then, by the definition of the class $\mathcal{U}_{k_1,k_2}(A)$ in (3.46) we have

$$1-(\alpha+\beta) \leq \mathsf{P}_A\left(D \in \mathcal{U}_{k_1,k_2}(\alpha,\beta)\right) = \sum_{B\in\mathcal{U}_{k_1,k_2}(\alpha,\beta)} \mathsf{P}_A(D=B).$$

Fix $B \in \mathcal{U}_{k_1,k_2}(\alpha,\beta)$, and let $\eta > 0$. First, we assume that $|B| \geq k_1$ and $|B^c| \geq k_2$. Then $\mathsf{P}_A(D=B)$ is upper bounded by $\mathrm{I}+\mathrm{II}$, where

$$\begin{aligned}
\mathrm{I} &= \mathsf{P}_A\left(\lambda^{A,B_1^*}(T) < \log(\frac{\eta}{\alpha}), D=B\right) + \mathsf{P}_A\left(\lambda^{A,B_2^*}(T) < \log(\frac{\eta}{\beta}), D=B\right)\\
\mathrm{II} &= \mathsf{P}_A\left(\lambda^{A,B_1^*}(T) \geq \log(\frac{\eta}{\alpha}), \lambda^{A,B_2^*}(T) \geq \log(\frac{\eta}{\beta}), D=B\right),
\end{aligned}$$

where the sets B_1^* and B_2^* are selected to satisfy the conditions in Case 1 of Lemma 3.5. Then, $|B\setminus B_1^*| \geq k_1$ and $|B_2^*\setminus B| \geq k_2$, and consequently

$$\mathsf{P}_{B_1^*}(D=B) \leq \alpha \ \text{ and } \ \mathsf{P}_{B_2^*}(D=B) \leq \beta.$$

Thus, by change of measure $\mathsf{P}_A \to \mathsf{P}_{B_1^*}$ and $\mathsf{P}_A \to \mathsf{P}_{B_2^*}$, we have

$$\mathsf{P}_A\left(\lambda^{A,B_i^*}(T) < \log\left(\frac{\eta}{\alpha}\right), D=B\right) \leq \eta, \ \text{ for } \ i=1,2$$

which shows that $\mathrm{I} \leq 2\eta$. Moreover, it is obvious that

$$\begin{aligned}
\mathrm{II} \leq& \epsilon^B_{\alpha,\beta} + \mathsf{P}_A(T \geq qL(A;\alpha,\beta),\ D=B), \quad \text{where}\\
\epsilon^B_{\alpha,\beta} :=& \mathsf{P}_A\left(T < qL(A;\alpha,\beta),\ \lambda^{A,B_1^*}(T) \geq \log\left(\frac{\eta}{\alpha}\right), \lambda^{A,B_2^*}(T) \geq \log\left(\frac{\eta}{\beta}\right)\right).
\end{aligned}$$

But by the construction of B_1^* and B_2^* we have

$$L(A;\alpha,\beta) \leq \ell_{\alpha,\beta} := \frac{|\log(\alpha)|}{\mathcal{I}^{A,B_1^*}} \bigvee \frac{|\log(\beta)|}{\mathcal{I}^{A,B_2^*}},$$

consequently

$$\epsilon^B_{\alpha,\beta} \leq \mathsf{P}_A\left(T < q\ell_{\alpha,\beta},\ \ \lambda^{A,B_1^*}(T) \geq \log\left(\frac{\eta}{\alpha}\right), \lambda^{A,B_2^*}(T) \geq \log\left(\frac{\eta}{\beta}\right)\right),$$

and from Lemma 3.13 it follows that $\epsilon^B_{\alpha,\beta}$ goes to 0 as $\alpha, \beta \to 0$.

Putting everything together, we have

$$\mathsf{P}_A(D = B) \leq 2\eta + \epsilon^B_{\alpha,\beta} + \mathsf{P}_A(T \geq qL(A;\alpha,\beta), D = B). \tag{3.47}$$

In a similar way we can show that equation (3.47) remains valid when $|B| < k_1$ or $|B^c| < k_2$. Thus summing over $B \in \mathcal{U}_{k_1,k_2}(A)$ we have

$$\mathsf{P}_A(D \in \mathcal{U}_{k_1,k_2}(A)) \leq 2Q\eta + \epsilon_{\alpha,\beta} + \mathsf{P}_A(T \geq qL(A;\alpha,\beta), D \in \mathcal{U}_{k_1,k_2}(A)),$$

where $Q = |\mathcal{U}_{k_1,k_2}(A)|$ is a constant, and $\epsilon_{\alpha,\beta} = \sum_{B\in\mathcal{U}_{k_1,k_2}(A)} \epsilon^B_{\alpha,\beta}$. Since each summand goes to 0, we have $\epsilon_{\alpha,\beta} \to 0$ as $\alpha, \beta \to 0$. Therefore,

$$\mathsf{P}_A(T \geq qL(A;\alpha,\beta)) \geq 1 - (\alpha + \beta) - 2Q\eta - \epsilon_{\alpha,\beta}$$

The proof is complete after taking the infimum over the class $\Delta_{k_1,k_2}(\alpha,\beta)$, letting $\alpha, \beta \to 0$ and letting $\eta \to 0$. □

3.10.5 Proof of Corollary 3.2

Proof. The error control for δ_0 follows by setting $\ell = 0$ in Lemma 3.4. The error control for the Intersection rule δ_I can be established by a simple modification of the proof of Lemma 3.4. If assumptions (3.11) and (3.12) hold, then from (3.26) it follows that for every $A \subset [J]$ we have

$$L_A(k_1, k_1, \alpha, \alpha) = \frac{|\log(\alpha)|}{k_1\mathcal{I}}.$$

Further, setting $\ell = 0$ for τ_0, and $k = 1$ for T_I in the first inequality of Lemma 3.1, we have as $b \to \infty$

$$\mathsf{E}_A\left[\tau_0(b,b)\right] \leq \frac{b}{k_1\mathcal{I}}(1 + o(1)), \quad \mathsf{E}_A\left[\tau_I(b,b)\right] \leq \frac{b}{\mathcal{I}}(1 + o(1)).$$

Thus, if b is selected as in the statement of the corollary, then the quantity $L_A(k_1, k_1, \alpha, \alpha)$ provides an asymptotic power bound for both $\mathsf{E}_A[\tau_0]$ and $\mathsf{E}_A[\tau_I]$. Thus, the proof is complete. □

3.10.6 Proof of Theorem 3.7

Proof. Since k_1, d are fixed, we write $n^*(\beta)$ and $\widehat{n}(\beta)$ for $n^*(k_1, k_1, \beta^d, \beta)$ and $\widehat{n}_{NP}(k_1, k_1, \beta^d, \beta)$ respectively for simplicity.

(i) Let us first focus on $n^*(\beta)$. By its definition (3.13), there exists some

$$D^*(\beta) \in \Delta_{fix}(n^*(\beta)) \cap \Delta_{k_1,k_1}(\beta^d, \beta).$$

Fix any $A \subset [J]$ such that $|A| = 2k_1 - 1$. Denote P the probability measure for data in all streams, and consider the simple versus simple testing problem (3.45) and the procedure $\widetilde{D}^*(\beta) := \begin{cases} 0 & \text{if } |D^*(\beta)| < k_1 \\ 1 & \text{if } |D^*(\beta)| \geq k_1 \end{cases}$. Then by definition of $D^*(\beta)$, we have

$$\begin{aligned} \mathsf{P}_\emptyset(\widetilde{D}^*(\beta) = 1) = \mathsf{P}_\emptyset(|D^*(\beta)| \geq k_1) \leq \alpha = \beta^d, \\ \mathsf{P}_A(\widetilde{D}^*(\beta) = 0) = \mathsf{P}_A(|D^*(\beta)| < k_1) \leq \beta, \end{aligned}$$

Then by the generalized Chernoff's Lemma 3.15,

$$\begin{aligned} \liminf_{\beta \to 0} \frac{1}{n^*(\beta)} \log(\beta) &\geq \liminf_{\beta \to 0} \frac{1}{n^*(\beta)} \log\left(\frac{1}{2}\mathsf{P}_\emptyset^{1/d}(\widetilde{D}^*(\beta) = 1) + \frac{1}{2}\mathsf{P}_A(\widetilde{D}^*(\beta) = 0)\right) \\ &\geq -\frac{\Phi^A(\widetilde{h}_d^A)}{d}. \end{aligned}$$

where $\widetilde{h}_d^A$ is a solution to $\Phi^A(z)/d = \Phi^A(z) - z$, and for any $z \in \mathbb{R}$

$$\begin{aligned} \Phi^A(z) &:= \sup_{\theta \in \mathbb{R}} \left\{ z\theta - \sum_{j \in A} \log\left(\mathsf{E}_0^j\left[e^{\theta \lambda^j(1)}\right]\right) \right\} \\ &= \sup_{\theta \in \mathbb{R}} \left\{ z\theta - |A| \log\left(\mathsf{E}_0^1\left[e^{\theta \lambda^1(1)}\right]\right) \right\} = |A| \Phi(\frac{z}{|A|}). \end{aligned}$$

Here, the second equality is due to homogeneity (3.27). By definition (3.29), $\Phi(h_d)/d = \Phi(h_d) - h_d$, which implies

$$\Phi^A(|A|h_d)/d = \Phi^A(|A|h_d) - (|A|h_d).$$

Thus $\widetilde{h}_d^A = |A|h_d$, and

$$\Phi^A(\widetilde{h}_d^A)/d = |A|\Phi(h_d)/d = \frac{2k_1 - 1}{d}\Phi(h_d).$$

which completes the proof of (i).

(ii) Let us now focus on $\widehat{n}(\beta)$. By definition, there exists $h_\beta \in \mathbb{R}$ such that

$$(\widehat{n}(\beta), \widehat{D}(\beta)) \in \Delta_{k_1,k_1}(\beta^d, \beta), \text{ where } \widehat{D}(\beta) := D_{NP}(\widehat{n}(\beta), h_\beta \mathbf{1}_J),$$

where $\mathbf{1}_J \in \mathbb{R}^J$ is a vector of all ones. Due to homogeneity (3.27), denote

$$p_\beta := \mathsf{P}_0^1(\widehat{D}^1(\beta) = 1) = \mathsf{P}_0^1\left(\frac{1}{\widehat{n}(\beta)}\lambda^1(\widehat{n}(\beta)) > h_\beta\right)$$
$$q_\beta := \mathsf{P}_1^1(\widehat{D}^1(\beta) = 0) = \mathsf{P}_1^1\left(\frac{1}{\widehat{n}(\beta)}\lambda^1(\widehat{n}(\beta)) \leq h_\beta\right)$$

For any $A \subset [J]$ such that $|A| = k_1 (= k_2)$,

$$\beta^d \geq \mathsf{P}_\emptyset\left(\cap_{j\in A}\{\tilde{D}(\alpha)^j = 1\}\right) = (p_\beta)^{k_1}, \;\; \beta \geq \mathsf{P}_{[J]}\left(\cap_{j\in A}\{\tilde{D}(\alpha)^j = 0\}\right) = (q_\beta)^{k_1},$$

which implies that

$$\frac{1}{\widehat{n}(\beta)}\frac{\log(\beta)}{k_1} \geq \frac{1}{\widehat{n}(\beta)}\log\left(\frac{1}{2}p_\beta^{1/d} + \frac{1}{2}q_\beta\right).$$

Then again by the generalized Chernoff's lemma 3.15, we have

$$\liminf_{\beta\to 0}\frac{\widehat{n}(\beta)}{|\log(\beta)|} = \frac{d}{k_1\Phi(h_d)}.$$

Further, the same argument shows that the equality is obtained with $h = h_d$, which completes the proof of (ii). □

3.11 Sequential multiple testing with composite hypotheses

In this section, we prove Theorem 3.8 in Section 3.6. We first establish a universal asymptotic lower bound on the expected sample size of procedures that control generalized familywise error rates under composite hypotheses (Subsec. 3.11.1). Then, we show that this lower bound is achieved by the Leap rule with the

adaptive log-likelihood statistics in (3.31) (Subsec. 3.11.2 and 3.11.3). Further, we demonstrate via numerical study that in the composite case, the Intersection rule (3.10) and the asymmetric Sum-Intersection rule (3.23) with the adaptive statistics fail to achieve asymptotic optimality (Subsec. 3.11.4). We conclude this section with discussions on the adaptive statistics and alternative local test statistics (Subsec. 3.11.5).

3.11.1 Lower bound on the expected sample size

Fix any $A \subset [J]$ and $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta}_A$.

Case 1: Assume *for now* that the infima in (3.33) are attained, i.e., there exists $\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}^1, \ldots, \widetilde{\theta}^J) \in \boldsymbol{\Theta}_{A^c}$ such that

$$\mathcal{I}_1^j(\theta^j) = I^j(\theta^j, \widetilde{\theta}^j) \text{ for any } j \in A,$$
$$\mathcal{I}_0^j(\theta^j) = I^j(\theta^j, \widetilde{\theta}^j) \text{ for any } j \in A^c.$$

Any procedure $(T, D) \in \Delta_{k_1,k_2}^{comp}(\alpha, \beta)$ controls the generalized familywise error rates below α and β when applied to the multiple testing problem with the following *simple* hypotheses for each stream:

$$\mathsf{H}_0^{j'} : \theta^{j'} = \theta^j \quad \text{versus} \quad \mathsf{H}_1^{j'} : \theta^{j'} = \widetilde{\theta}^j, \quad j \in [J],$$

where we write $\theta^{j'}$ for the generic local parameter in j-th stream to distinguish it from the j-th component of $\boldsymbol{\theta}$.

Then, under assumptions (3.32) and (3.34), by Theorem 3.6 we have

$$\liminf_{\alpha \wedge \beta \to 0} N_{A,\boldsymbol{\theta}}^*(k_1, k_2, \alpha, \beta) / L_{A,\boldsymbol{\theta}}(k_1, k_2, \alpha, \beta) \geq 1, \tag{3.48}$$

where

$$\begin{aligned}
L_{A,\boldsymbol{\theta}}(k_1, k_2, \alpha, \beta) &:= \min\left\{ \min_{0 \leq \ell < k_1} \widehat{L}_{A,\boldsymbol{\theta}}(\ell; \alpha, \beta) \, , \, \min_{0 \leq \ell < k_2} \check{L}_{A,\boldsymbol{\theta}}(\ell; \alpha, \beta) \right\}, \\
\widehat{L}_{A,\boldsymbol{\theta}}(\ell; \alpha, \beta) &:= \max\left\{ \frac{|\log(\alpha)|}{\mathcal{D}_1(A, \boldsymbol{\theta}; 1, k_1 - \ell)} \, , \, \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, \boldsymbol{\theta}; \ell + 1, \ell + k_2)} \right\}, \\
\check{L}_{A,\boldsymbol{\theta}}(\ell; \alpha, \beta) &:= \max\left\{ \frac{|\log(\alpha)|}{\mathcal{D}_1(A, \boldsymbol{\theta}; \ell + 1, \ell + k_1)} \, , \, \frac{|\log(\beta)|}{\mathcal{D}_0(A^c, \boldsymbol{\theta}; 1, k_2 - \ell)} \right\}, \\
\mathcal{D}_1(A, \boldsymbol{\theta}; \ell, u) &= \sum_{j=\ell}^{u} \mathcal{I}_1^{(j)}(A, \boldsymbol{\theta}), \quad \mathcal{D}_0(A^c, \boldsymbol{\theta}; \ell, u) = \sum_{j=\ell}^{u} \mathcal{I}_0^{(j)}(A^c, \boldsymbol{\theta}),
\end{aligned} \tag{3.49}$$

and

$$\mathcal{I}_1^{(1)}(A,\boldsymbol{\theta}) \le \ldots \le \mathcal{I}_1^{(|A|)}(A,\boldsymbol{\theta})$$

is the increasingly ordered sequence of $\{\mathcal{I}_1^j(\theta^j), j \in A\}$, and

$$\mathcal{I}_0^{(1)}(A^c,\boldsymbol{\theta}) \le \ldots \le \mathcal{I}_0^{(|A^c|)}(A^c,\boldsymbol{\theta})$$

is the increasingly ordered sequence of $\{\mathcal{I}_0^j(\theta^j), j \in A^c\}$. As before, the convention is that

$$\mathcal{I}_1^{(k)}(A,\boldsymbol{\theta}) = \infty \text{ if } k > |A|, \quad \mathcal{I}_0^{(k)}(A^c,\boldsymbol{\theta}) = \infty \text{ if } k > |A^c|.$$

Case 2: In general, the infima in (3.33) are not attained. However, under the separability assumption (3.34), for any $\epsilon > 0$ there exists $\widetilde{\boldsymbol{\theta}}_\epsilon = (\widetilde{\theta}_\epsilon^1, \ldots, \widetilde{\theta}_\epsilon^J) \in \boldsymbol{\Theta}_{A^c}$ such that

$$I^j(\theta^j, \widetilde{\theta}_\epsilon^j) \le (1+\epsilon)\,\mathcal{I}_1^j(\theta^j) \text{ for any } j \in A,$$
$$I^j(\theta^j, \widetilde{\theta}_\epsilon^j) \le (1+\epsilon)\,\mathcal{I}_0^j(\theta^j) \text{ for any } j \in A^c.$$

Applying again Theorem 3.6 to the following multiple testing problem with simple hypotheses:

$$\mathsf{H}_0^{j'} : \theta^{j'} = \theta^j \quad \text{versus} \quad \mathsf{H}_1^{j'} : \theta^{j'} = \widetilde{\theta}_\epsilon^j, \quad j \in [J],$$

we have

$$\liminf \; N^*_{A,\boldsymbol{\theta}}(k_1,k_2,\alpha,\beta)/L_{A,\boldsymbol{\theta}}(k_1,k_2,\alpha,\beta) \ge 1/(1+\epsilon).$$

Since ϵ is arbitrary, (3.48) still holds.

Above discussions leads to the following theorem.

Theorem 3.9. *If* (3.32) *and* (3.34) *hold, then* (3.48) *holds for every* $A \subset [J]$ *and* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_A$.

3.11.2 Error control of the Leap rule with adaptive log-likelihood ratios

We start with the following observation.

Lemma 3.8. *Fix $A \subset [J]$, $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J) \in \Theta_A$. For each $j \in [J]$,*

$$L_n^j := \exp\left(\ell_*^j(n) - \ell^j(n, \theta^j)\right), \quad n \in \mathbb{N}$$

is an $\{\mathcal{F}_n\}$-martingale under $\mathsf{P}_{A,\boldsymbol{\theta}}$ with expectation 1.

Proof. By definition,

$$L_n^j = L_{n-1}^j \cdot \frac{p^j_{\widehat{\theta}^j_{n-1}}\left(X^j(n)|\mathcal{F}^j_{n-1}\right)}{p^j_{\theta^j}\left(X^j(n)|\mathcal{F}^j_{n-1}\right)} \ .$$

Clearly, $L_n^j \in \mathcal{F}_n$ for any $n \in \mathbb{N}$. Further, since $\widehat{\theta}^j_{n-1} \in \mathcal{F}_{n-1}$,

$$\mathsf{E}_{A,\boldsymbol{\theta}}\left[\frac{p^j_{\widehat{\theta}^j_{n-1}}\left(X^j(n)|\mathcal{F}^j_{n-1}\right)}{p^j_{\theta^j}\left(X^j(n)|\mathcal{F}^j_{n-1}\right)}\middle| \mathcal{F}_{n-1}\right] = \int \frac{p^j_{\widehat{\theta}^j_{n-1}}\left(z|\mathcal{F}^j_{n-1}\right)}{p^j_{\theta^j}\left(z|\mathcal{F}^j_{n-1}\right)} p^j_{\theta^j}\left(z|\mathcal{F}^j_{n-1}\right) = 1,$$

which implies $\mathsf{E}_{A,\boldsymbol{\theta}}[L_n^j|\mathcal{F}_{n-1}] = L_{n-1}^j$. Further, since $\widehat{\boldsymbol{\theta}}_0$ is deterministic, $\mathsf{E}_{A,\boldsymbol{\theta}}[L_1^j] = 1$, which completes the proof. □

By Lemma 3.8 and due to independence across streams, for any subset $M \subset [J]$, there exists a probability measure $Q_{A,\boldsymbol{\theta},M}$ such that for any $n \in \mathbb{N}$,

$$\frac{dQ_{A,\boldsymbol{\theta},M}}{d\mathsf{P}_{A,\boldsymbol{\theta}}}(\mathcal{F}_n) \ = \ \prod_{j\in M} \exp\left(\ell_*^j(n) - \ell^j(n,\theta^j)\right). \tag{3.50}$$

Next, we establish the error control of the Leap rule with adaptive log-likelihood ratios. The proof is almost identical to Theorem 3.5.

Theorem 3.10. *Assume* (3.33) *and* (3.34) *hold. For any $\alpha, \beta \in (0,1)$ we have that the Leap rule $\delta_L^*(a,b) \in \Delta^{comp}_{k_1,k_2}(\alpha,\beta)$ when the thresholds are selected as follows:*

$$a = |\log(\beta)| + \log(2^{k_2} C_{k_2}^J), \quad b = |\log(\alpha)| + \log(2^{k_1} C_{k_1}^J).$$

Proof. Just as the proof of Theorem 3.5 in Section 3.10.1 follows directly from Lemma 3.4, by exactly the same argument, the above result follows from the next Lemma. □

Lemma 3.9. *Assume* (3.33) *and* (3.34) *hold. Fix $A \subset [J]$, $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta}_A$. Let $B_1 \subset A^c$ with $|B_1| = k_1$, and $B_2 \subset A$ with $|B_2| = k_2$.*

(i) Fix any $0 \le \ell < k_1$. For any event $\Gamma \in \mathcal{F}_{\widehat{\tau}_\ell}$, we have

$$\mathsf{P}_{A,\boldsymbol{\theta}}(B_1 \subset \widehat{D}^*_\ell) \le C_\ell^{k_1} e^{-b}, \quad \mathsf{P}_{A,\boldsymbol{\theta}}(B_2 \subset (\widehat{D}^*_\ell)^c, \Gamma) \le e^{-a} Q_{A,\boldsymbol{\theta},B_2}(\Gamma).$$

(ii) Fix any $0 \le \ell < k_2$. For any event $\Gamma \in \mathcal{F}_{\check{\tau}_\ell}$, we have

$$\mathsf{P}_{A,\boldsymbol{\theta}}(B_1 \subset \check{D}^*_\ell, \Gamma) \le e^{-b} Q_{A,\boldsymbol{\theta},B_1}(\Gamma), \quad \mathsf{P}_{A,\boldsymbol{\theta}}(B_2 \subset (\check{D}^*_\ell)^c) \le C_\ell^{k_2} e^{-a}.$$

Proof. The proof is similar to that of Lemma 3.4. We only indicate the differences by working out the first inequality in (i).

As in the proof of Lemma 3.4, by definition, $\widehat{D}^*_\ell$ rejects the nulls in the ℓ streams with the least significant non-positive LLR, in addition to the nulls in the streams with positive LLR. Thus,

$$\{B_1 \subset \widehat{D}^*_\ell\} \subset \bigcup_{M \subset B_1, |M| = k_1 - \ell} \Pi_M \ , \quad \text{where } \Pi_M := \{\lambda^j_*(\widehat{\tau}_\ell) > 0 \ \ \forall j \ \in M\},$$

and by Boole's inequality, it suffices to show that $\mathsf{P}_{A,\boldsymbol{\theta}}(\Pi_M) \le e^{-b}$ for any $M \subset B_1$ with $|M| = k_1 - \ell$.

By definition, for any $j \in M \subset B_1 \subset A^c$, since $\theta^j \in \Theta^j_0$,

$$\ell^j_0(n) \ge \ell^j(n, \theta^j) \quad \text{for any } n \in \mathbb{N}.$$

Then, by the definition of the adaptive log-likelihood ratio statistics (3.31), we have

$$\Pi_M \ \subset \ \left\{ \sum_{j \in M} \left(\ell^j_*(\widehat{\tau}_\ell) - \ell^j_0(\widehat{\tau}_\ell) \right) \ge b \right\} \ \subset \ \left\{ \sum_{j \in M} \left(\ell^j_*(\widehat{\tau}_\ell) - \ell^j(\widehat{\tau}_\ell, \theta^j) \right) \ge b \right\}.$$

By the above observation, the definition of $Q_{A,\boldsymbol{\theta},M}$ (3.50), and likelihood ratio identity, on the event Π_M,

$$\frac{dQ_{A,\boldsymbol{\theta},M}}{d\mathsf{P}_{A,\boldsymbol{\theta}}}(\mathcal{F}_{\widehat{\tau}_\ell}) \ge e^b,$$

and the proof is complete by changing the measure from $\mathsf{P}_{A,\boldsymbol{\theta}}$ to $Q_{A,\boldsymbol{\theta},M}$. □

3.11.3 Asymptotic optimality of the Leap rule with adaptive log-likelihood ratios

The asymptotic optimality follows after we establish an upper bound on the expected sample size of the Leap rule. The following result is similar to Lemma 3.1.

Lemma 3.10. *Assume* (3.34) *and* (3.35) *hold. For any* $A \subset [J]$ *and* $\mathsf{\theta} \in \Theta_A$*, as* $a, b \to \infty$*,*

$$\mathsf{E}_{A,\mathsf{\theta}}[\widehat{\tau}_\ell] \le \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A,\mathsf{\theta};1,k_1-\ell)} \, , \, \frac{a(1+o(1))}{\mathcal{D}_0(A^c,\mathsf{\theta};\ell+1,\ell+k_2)} \right\}, \; 0 \le \ell < k_1,$$
$$\mathsf{E}_{A,\mathsf{\theta}}[\widecheck{\tau}_\ell] \le \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A,\mathsf{\theta};\ell+1,\ell+k_1)} \, , \, \frac{a(1+o(1))}{\mathcal{D}_0(A^c,\mathsf{\theta};1,k_2-\ell)} \right\}, \; 0 \le \ell < k_2.$$

where the denominators are defined in (3.49)*.*

Proof. Under the assumption (3.35), the proof is the same as that for Lemma 3.1 in Subsection 3.10.2. □

Now Theorem 3.8 follows from Theorem 3.9, Lemma 3.10 and Lemma 3.10.

3.11.4 Simulations for composite case

We consider a "homogeneous" multiple testing problem on the normal means with known variance. Specifically, we assume that for each $j \in [J]$, the sequence of observations in the j-th stream, $\{X^j(n) : n \in \mathbb{N}\}$, are i.i.d. with common distribution $\mathcal{N}(\theta^j, 1)$, and for a given constant $\mu > 0$, that does not depend on j, we want to test

$$\mathsf{H}_0^j \; : \theta^j \le 0 \quad \text{versus} \quad \mathsf{H}_1^j \; : \theta^j \ge \mu. \tag{3.51}$$

Instead of the Lebegure measure on the real line, we chose $\mathcal{N}(0,1)$ as our reference measure. Then, for each $j \in [J]$ we have

$$\ell^j(n, \theta^j) = n\left(\theta^j \, \overline{X}^j(n) - \frac{1}{2}(\theta^j)^2\right), \quad \text{where } \overline{X}^j(n) := \frac{1}{n}\sum_{i=1}^{n} X^j(i). \tag{3.52}$$

Further, for any $\theta^j, \widetilde{\theta}^j$, we have $I^j(\theta^j, \widetilde{\theta}^j) = \frac{1}{2}(\theta^j - \widetilde{\theta}^j)^2$, and

$$I_0^j(\theta^j) = \frac{1}{2}(\theta^j - \mu)^2 \text{ for } \theta^j \le 0, \qquad I_1^j(\theta^j) = \frac{1}{2}(\theta^j)^2 \text{ for } \theta^j \ge \mu.$$

Clearly, the null and the alternative hypotheses are separated in the sense of (3.34). Further, the condition (3.32) is satisfied due to [31].

The adaptive log-likelihood process (3.30) for the j-th stream in this context takes the following form: $\ell_0^j = 0$, and for $n \geq 1$,

$$\ell_*^j(n) = \sum_{i=1}^{n} \left(X^j(i)\, \widehat{\theta}_{i-1}^j - \frac{1}{2}(\widehat{\theta}_{i-1}^j)^2 \right). \tag{3.53}$$

If we choose to use the maximum likelihood estimators $\{\widehat{\boldsymbol{\theta}}_n\}$ in above definition, i.e., $\widehat{\theta}_n^j = \overline{X}^j(n)$, the one-sided complete convergence condition (3.35) is established in [71] (Page 278-279). Thus by Theorem 3.8, the Leap rule is asymptotically optimal in this setup.

To distinguish from the simulations in the simple versus simple setup, we refer to the Leap rule with adaptive statistics as "Leap*" rule. We will compare the Leap* rule with the following procedures:

1. *Asymmetric Sum-Intersection* rule*: replace the log-likelihood ratio statistics $\lambda^j(n)$, in the definition of the asymmetric Sum-Intersection rule (3.23), by the adaptive version $\lambda_*^j(n)$ (3.30).

2. *Intersection* rule*: replace the log-likelihood ratio statistics $\lambda^j(n)$, in the definition of the Intersection rule (3.10), by the adaptive version $\lambda_*^j(n)$ (3.30).

3. *MNP rule*: for a fixed-sample size n, in each stream, we run the Neyman-Pearson rule with the same threshold $h > 0$, which is the most powerful test for each stream due to monotone likelihood ratio property. Formally,

$$\delta_{NP}(n,h) := (n, D_{NP}(n,h)), \quad D_{NP}(n,h) := \{j \in [J] : \overline{X}^j(n) > h\},$$

For simulation purpose, we assume that the tolerance on the two types of mistakes are the same in the sense that (3.12) holds. As in Section 3.6, we denote the true parameter as $(A, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta}_A$.

Thresholds selection via simulation

For each $j \in [J]$ and $\theta^j \leq 0$ the distribution of $\{\lambda_*^j(n) : n \in \mathbb{N}\}$ under $\mathsf{P}^j_{\mu-\theta^j}$ is the same as the distribution of $\{-\lambda_*^j(n) : n \in \mathbb{N}\}$ under $\mathsf{P}^j_{\theta^j}$. Since (3.12) holds, we should equate the thresholds a and b in the Leap* rule. Further, we only need to focus on the generalized familywise error rate of Type I.

For a fixed parameter a $(= b)$, we use simulation to find out the maximal probability of the Leap* rule committing k_1 false positive mistakes, i.e.

$$\max_{(A,\boldsymbol{\theta}): A \subset [J], \boldsymbol{\theta} \in \boldsymbol{\Theta}_A} \mathsf{P}_{A,\boldsymbol{\theta}}(|D_L^* \setminus A| \geq k_1).$$

Then we try different values for a and select the one for which the above quantity is equal to α. Note that the maximum is over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. However, for $\theta^j \leq \widetilde{\theta^j}$, $\{\lambda_*^j(n) : n \in \mathbb{N}\}$ under $\mathsf{P}_{\widetilde{\theta^j}}$ is stochastically larger than $\{\lambda_*^j(n) : n \in \mathbb{N}\}$ under P_{θ^j}, in the sense that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$\mathsf{P}^j_{\theta^j}(\lambda_*^j(n) \leq x) \;\geq\; \mathsf{P}^j_{\widetilde{\theta^j}}(\lambda_*^j(n) \leq x).$$

As a result, the maximal probability is achieved by the boundary cases, i.e., $\boldsymbol{\theta} \in \{0, \mu\}^J$.

The same discussion applies to the other two sequential procedures. For the MNP rule, (3.12) implies that $h = \frac{1}{2}\mu$, and for a fixed n, the maximal probability of making k_1 false positives is also achieved by $\boldsymbol{\theta} \in \{0, \mu\}^J$.

Practical considerations

The first few estimators of $\boldsymbol{\theta}$ will typically be quite noisy, since they are estimated based on only a few observations. However, from (3.30) or (3.53) we observe that their effect will persist. Thus, in practice it is preferable to take an initial sample of fixed size, say n_0, and use these observations *only* to obtain good initial estimates of the unknown parameter.

Specifically, we assume that for each $j \in [J]$, $X^j(-n_0), \ldots, X^j(-1)$ are i.i.d. with distribution $\mathcal{N}(\theta^j, 1)$, and we define for $n \geq 0$ the following maximum likelihood estimator

$$\widehat{\theta}_n^j := \frac{\sum_{i=-n_0}^{-1} X^j(i) + \sum_{i=1}^{n} X^j(i)}{n_0 + n},$$

which includes the initial samples. The definitions of the log-likelihood process (3.52) and the adaptive log-likehood process (3.53) remain unchanged. By taking an initial sample of fixed size, the asymptotic expected sample size of the Leap* rule is not affected. Further, if we enlarge the σ-field by including the initial samples, i.e.,

$$\widetilde{\mathcal{F}}_n := \mathcal{F}_n \;\vee\; \sigma\left(X^j(i) : j \in [J], i \in \{-n_0, \ldots, -1\}\right),$$

then the key Lemma 3.8, used to establish the error control of Leap* rule, still holds. Thus, taking an initial sample does not affect the asymptotic optimality of the Leap* rule.

Simulation results

We consider the problem (3.51) with $J = 20$, $\mu = 0.2$, $k_1 = k_2 = 2$ and the initial sample size $n_0 = 10$. Based on the previous discussion, we set $a = b$ for the sequential methods. For a fixed threshold a, we use

simulation to find out the maximal probability (over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$) of committing k_1 false positives (Err), and the expected sample size (ESS) under a particular $\mathsf{P}_{A,\boldsymbol{\theta}}$, where $A = \{1, \ldots, 10\}$ and

$$\boldsymbol{\theta} = (\theta^1, \ldots, \theta^J), \quad \theta^j = \begin{cases} 0.7 & \text{if } j = 1, \ldots, 10 \\ -0.3 & \text{if } j = 11, \ldots, 19 \\ 0 & \text{if } j = 20 \end{cases} . \tag{3.54}$$

For the MNP rule, we set $h = \frac{1}{2}\mu$, and use simulation to find out the maximal probability of committing k_1 false positives for each fixed $n \in \mathbb{N}$. The results are shown in Figure 3.10.

From Figure 3.10, we observe that the other procedures have a different "slope" compared to the asymptotically optimal Leap* rule, which indicates that they fail to be asymptotically optimal. Further, since sequential methods are adaptive to the true $\boldsymbol{\theta}$, the gains over fixed-sample size procedures increase as $\boldsymbol{\theta}$ is farther from the boundary cases.

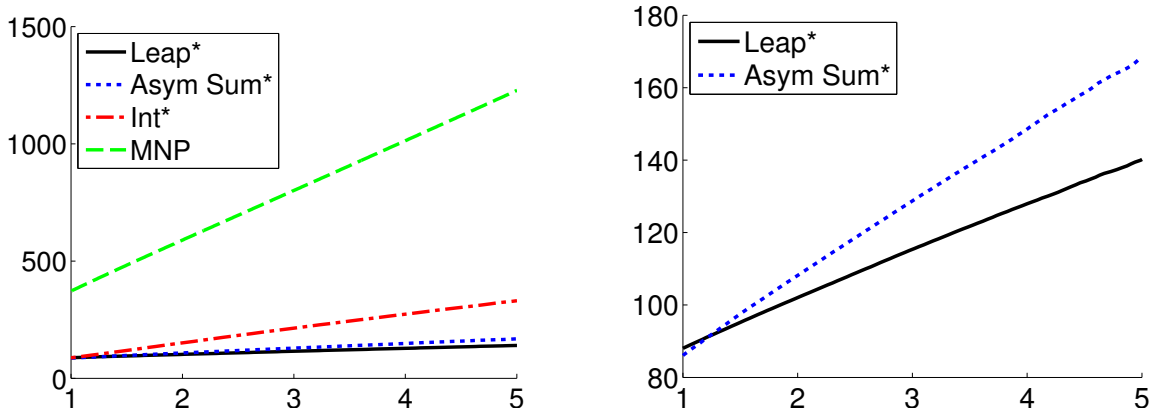


Figure 3.10: The testing problem (3.51) with $J = 20, \mu = 0.2, k_1 = k_2 = 2$ and the initial sample size $n_0 = 10$. The x-axis in both graphs is $|\log_{10}(\text{Err})|$. The y-axis is the corresponding ESS under $\boldsymbol{\theta}$ given by (3.54). The second figure plots two of the lines in the first figure. Note that for the sequential procedures, the initial sample size n_0 is added to the ESS.

3.11.5 Discussion on the local test statistics

When there is only one stream (i.e. $J = 1$), the adaptive log-likelihood ratio statistic (3.31) was first proposed in [58] in the context of power one tests, and later extended by [54] to sequential multi-hypothesis testing. There are two other popular choices for the local test statistics in the case of composite hypotheses.

The first one is to follow the approach suggested by Wald [75] and replace $\lambda^j(n)$ in the Leap rule (3.24)

by the following *mixture* log-likelihood ratio statistic:

$$\log\left(\frac{\int_{\Theta_1^j}\,\exp\left(\ell(n,\theta^j)\right)\;\omega_1^j(d\theta^j)}{\int_{\Theta_0^j}\,\exp\left(\ell(n,\theta^j)\right)\;\omega_0^j(d\theta^j)}\right),$$

where ω_0^j, ω_1^j are two probability measures on Θ_0^j and Θ_1^j respectively. The second is to replace $\lambda^j(n)$ in the Leap rule (3.24) by the *generalized* log-likelihood ratio (GLR) statistic $\widehat{\ell}_1^j(n) - \widehat{\ell}_0^j(n)$. When there is only one stream (i.e. $J = 1$), the corresponding sequential test has been studied in [43] for one-parameter exponential family, in [15] for multi-parameter exponential family, and in [38] for separate families of hypotheses.

We have chosen the adaptive log-likelihood ratio statistics (3.31) in this Chapter mainly because they allow for explicit and universal error control. Indeed, with this choice of statistics, the upper bounds on the error probabilities rely on a change-of-measure argument, in view of Lemma 3.8, whereas this argument breaks down when we use GLR or mixture statistics.

3.12 Sequential testing of two composite hypotheses in exponential family

In this section, we show that (3.35) holds if each stream has i.i.d. observations from an exponential family distribution, both the null and alternative parameter spaces are compact, and the maximal likelihood estimator is used in the adaptive log-likelihood statistics (3.31). Note that (3.35) is a condition on each *individual stream*, thus in this section we drop the superscript j.

Let $\{X_n : n \in \mathbb{N}\}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ with common density

$$p_\theta(x) = \exp\left(\theta^T x - b(\theta)\right)$$

with respect to some measure ν, where superscript T means transpose. We assume that the natural parameter space

$$\Theta := \{\theta \in \mathbb{R}^d : \int p_\theta(x)\nu(dx) < \infty\}$$

is an open subset of $\mathbb{R}^d$. For any $\theta, \widetilde{\theta} \in \Theta$, the Kullback-Leibler divergence between p_θ and $p_{\widetilde{\theta}}$ is denoted by

$$I(\theta, \widetilde{\theta}) := \mathsf{E}_\theta\left[\log\frac{p_\theta(X_1)}{p_{\widetilde{\theta}}(X_1)}\right] = (\theta - \widetilde{\theta})^T \nabla b(\theta) - (b(\theta) - b(\widetilde{\theta})),$$

where ∇ stands for the gradient. We denote by $\{\ell(n,\theta) : n \in \mathbb{N}\}$ the log-likelihood process:

$$\ell(n,\theta) := \sum_{i=1}^{n} \log p_\theta(X_i) = \sum_{i=1}^{n} (\theta^T X_i - b(\theta)) \quad \text{for } n \in \mathbb{N}.$$

We assume that Θ_0, Θ_1 are two *disjoint, compact* subsets of Θ, and denote by

$$\widehat{\theta}_n := \arg \max_{\theta \in \Theta_0 \cup \Theta_1} \ell(n,\theta)$$

the maximum likelihood estimator based on the data up to time n over the set $\Theta_0 \cup \Theta_1$. Picking any deterministic $\widehat{\theta}_0 \in \Theta$, we define

$$\ell_*(n) := \sum_{i=1}^{n} \log p_{\widehat{\theta}_{i-1}}(X_i) = \sum_{i=1}^{n} (\widehat{\theta}_{i-1}^T X_i - b(\widehat{\theta}_{i-1})) \quad \text{for } n \in \mathbb{N}.$$

The main result of this subsection is summarized in the following theorem.

Theorem 3.11. *Let $\theta \in \Theta_1$ and set $I(\theta) := \inf_{\theta_0 \in \Theta_0} I(\theta, \theta_0)$. Then, for any $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} \mathsf{P}_\theta \left(\frac{\ell_*(n) - \ell_0(n)}{n} - I(\theta) < \epsilon \right) < \infty,$$

where $\ell_0(n) := \sup_{\theta_0 \in \Theta_0} \ell(n, \theta_0)$.

Proof. Observe that for any $\theta_0 \in \Theta_0$,

$$\ell_*(n) - \ell(n,\theta_0) = \ell_*(n) - \ell(n,\theta) + \ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0) + nI(\theta,\theta_0),$$

which implies that

$$\begin{aligned} \ell_*(n) - \ell_0(n) &= \ell_*(n) - \ell(n,\theta) + \inf_{\theta_0 \in \Theta_0} (\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0) + nI(\theta,\theta_0)) \\ &\geq \ell_*(n) - \ell(n,\theta) + \inf_{\theta_0 \in \Theta_0} (\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0)) + nI(\theta). \end{aligned}$$

As a result, it suffices to show that

$$\frac{1}{n}(\ell_*(n) - \ell(n,\theta)) \xrightarrow[n\to\infty]{\mathsf{P}_\theta \text{ completely}} 0, \tag{3.55}$$

$$\frac{1}{n} \inf_{\theta_0 \in \Theta_0} (\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0)) \xrightarrow[n\to\infty]{\mathsf{P}_\theta \text{ completely}} 0, \tag{3.56}$$

which are the content of the next two lemmas. □

Remark 3.6. *The sequence in* (3.55) *concerns the behavior of the maximal likelihood estimator for the exponential family distribution, while the sequence in* (3.56) *concerns the uniform behavior over* Θ_0.

Lemma 3.11. *For any* $\theta \in \Theta$, *as* $n \to \infty$, $\frac{1}{n}(\ell_*(n) - \ell(n,\theta))$ *converges completely to zero under* P_θ.

Proof. Since Θ_0 and Θ_1 are compact, there exists $K > 0$ such that

$$\max\{\|\widetilde{\theta}\|,\ I(\theta,\widetilde{\theta})\} < K \quad \text{for any } \widetilde{\theta} \in \Theta_0 \cup \Theta_1,$$

where we use $\|\cdot\|$ to denote the Euclidean distance.

Observe that $\frac{1}{n}(\ell_*(n) - \ell(n,\theta)) = \frac{1}{n}M_n - \frac{1}{n}R_n$, where

$$M_n := \ell_*(n) - \ell(n,\theta) + \sum_{i=1}^{n} I(\theta,\widehat{\theta}_{i-1}) = \sum_{i=1}^{n} (\widehat{\theta}_{i-1} - \theta)^T (X_i - \nabla b(\theta)),$$
$$R_n := \sum_{i=1}^{n} I(\theta,\widehat{\theta}_{i-1})$$

Denote $\mathcal{F}_n := \sigma(X_1,\ldots,X_n)$ the σ-field generated by the first n observations. Then $\{M_n : n \in \mathbb{N}\}$ is an $\{\mathcal{F}_n\}$-martingale, since $\mathsf{E}[X_1] = \nabla b(\theta)$ due to the property of the exponential family and $\widehat{\theta}_{n-1} \in \mathcal{F}_{n-1}$. Further, the martingale difference sequence $\{(\widehat{\theta}_{i-1} - \theta)^T (X_i - \nabla b(\theta)) : i \in \mathbb{N}\}$ is bounded in L^p for any $p > 2$. Indeed, by Cauchy-Schwarz inequality,

$$\sup_{i\in\mathbb{N}} \mathsf{E}|(\widehat{\theta}_{i-1} - \theta)^T (X_i - \nabla b(\theta))|^p \le (2K)^p \mathsf{E}\|X_1 - \nabla b(\theta)\|^p < \infty.$$

Then by [68], we conclude $\frac{1}{n}M_n$ converges completely to zero under P_θ.

It remains to show that $\frac{1}{n}R_n$ converges completely to zero under P_θ. Fix any $\epsilon > 0$. Since $I(\theta,\widetilde{\theta})$ is continuous in $\widetilde{\theta}$, there exists $\delta > 0$ such that if $\|\widetilde{\theta} - \theta\| \le \delta$, $I(\theta,\widetilde{\theta}) \le \epsilon/2$. Define three random times

$$\eta_1 := \sup\{n \in \mathbb{N} : |\frac{1}{n}R_n| > \epsilon\},$$
$$\eta_2 := \sup\{n \in \mathbb{N} : |I(\theta,\widehat{\theta}_n)| > \epsilon/2\}, \quad \eta_3 := \sup\{n \in \mathbb{N} : \|\widehat{\theta}_n - \theta\| > \delta\}$$

By Theorem 5.1 in [54], there exist constant c_1 and c_2 such that $\mathsf{P}_\theta(\eta_3 > n) \le c_1 \exp(-c_2 n)$ for any $n \in \mathbb{N}$. In particular,

$$\mathsf{E}_\theta[\eta_3] < \infty.$$

Clearly, $\eta_2 \leq \eta_3$, which implies that $\mathsf{E}_\theta[\eta_2] < \infty$. We next show that $\eta_1 \leq 2\epsilon K \eta_2$. Indeed, for $n \geq 2K\eta_2/\epsilon$,

$$|\frac{1}{n}R_n| \leq \frac{1}{n}\left(\sum_{i=1}^{\eta_2} I(\theta, \widehat{\theta}_{i-1}) + \sum_{i=\eta_2+1}^{n} I(\theta, \widehat{\theta}_{i-1})\right) \leq \frac{K\eta_2 + \epsilon/2 * n}{n} \leq \epsilon.$$

Thus $\mathsf{E}_\theta[\eta_1] < \infty$, which implies $\frac{1}{n}R_n$ converges to zero quickly. (See Chapter 2.4.3 in [71] for formal definition of quick convergence.) Due to Lemma 2.4.1 in [71], quick convergence implies complete convergence, and thus $\frac{1}{n}R_n$ converges to zero completely. □

Lemma 3.12. *Assume the conditions in Theorem 3.11 hold. Then*

$$\frac{1}{n}\inf_{\theta_0 \in \Theta_0}(\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0)) \xrightarrow[n\to\infty]{\mathsf{P}_\theta \text{ completely}} 0.$$

Proof. By definition, we have

$$\begin{aligned}\frac{1}{n}\inf_{\theta_0 \in \Theta_0}(\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0)) &= \frac{1}{n}\inf_{\theta_0 \in \Theta_0}\sum_{i=1}^{n}(\theta-\theta_0)^T(X_i - \nabla b(\theta)) \\ &= \inf_{\theta_0 \in \Theta_0}(\theta-\theta_0)^T(\frac{1}{n}\sum_{i=1}^{n}(X_i - \nabla b(\theta))).\end{aligned}$$

Denote θ_j, $\theta_{0,j}$, $X_{i,j}$ and $\nabla_j b(\theta)$ the j^{th} dimension of the $\mathbb{R}^d$ vectors θ, θ_0, X_i and $\nabla b(\theta)$. Since Θ_0, Θ_1 is compact, there exists $K > 0$ such that

$$|\theta_j|, |\theta_{0,j}| \leq K, \text{ for any } 1 \leq j \leq d, \ \ \theta_0 \in \Theta_0.$$

By triangle inequality,

$$\left|\frac{1}{n}\inf_{\theta_0 \in \Theta_0}(\ell(n,\theta) - \ell(n,\theta_0) - nI(\theta,\theta_0))\right| \leq \ \ 2K\sum_{j=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{d}(X_{i,j} - \nabla_j b(\theta))\right|.$$

But for each $1 \leq j \leq d$, since $\mathsf{E}_\theta[X_{i,j}^2] < \infty$, by [31],

$$\frac{1}{n}\sum_{i=1}^{d}(X_{i,j} - \nabla_j b(\theta)) \xrightarrow[n\to\infty]{\mathsf{P}_\theta \text{ completely}} 0,$$

which completes the proof. □

3.13 Two renewal-type lemmas

In this section, we present two renewal-type lemmas about general discrete stochastic process, which may be of independent interest.

Lemma 3.13. *Let $\{\xi_i(n) : n \in \mathbb{N}\}$ $(i = 1, 2)$ be two stochastic processes on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Suppose that for some positive μ_1, μ_2,*

$$\mathsf{P}\left(\lim_{n\to\infty} \frac{1}{n}\xi_i(n) = \mu_i\right) = 1 \quad \text{for } i = 1, 2.$$

Let c be a fixed constant. Then for any random time T, and any $q \in (0, 1)$,

$$\lim_{b\to\infty} \mathsf{P}\left(T \le q\frac{b}{\mu_1},\ \xi_1(T) \ge b + c\right) = 0, \tag{3.57}$$

$$\lim_{a,\,b\to\infty} \mathsf{P}\left(T \le q(\frac{a}{\mu_1} \bigvee \frac{b}{\mu_2}),\ \xi_1(T) \ge a + c,\ \xi_2(T) \ge b + c\right) = 0. \tag{3.58}$$

Proof. Since c is fixed, we assume $c = 0$ without loss of generality. Denote $N_b = \lfloor q\frac{b}{\mu_1} \rfloor$, and $\epsilon_q = \frac{1}{q} - 1 > 0$. Notice that $\mathsf{P}(T \le q\frac{b}{\mu_1},\ \xi_1(T) \ge b)$ is upper bounded by

$$\mathsf{P}\left(\max_{1\le n\le N_b} \xi_1(n) \ge b\right) \ \le\ \mathsf{P}\left(\frac{1}{N_b} \max_{1\le n\le N_b} \xi_1(n) \ge (1 + \epsilon_q)\mu_1\right) \to 0$$

where the convergence follows directly from Lemma A.1 of [25]. Thus the proof of (3.57) is compelte.

For the second part, assume (3.58) doesn't hold. Then there exists some $\epsilon > 0$, and a sequence (a_n, b_n) with $a_n \to \infty, b_n \to \infty$ such that

$$p_n := \mathsf{P}\left(T \le q(\frac{a_n}{\mu_1} \bigvee \frac{b_n}{\mu_2}),\ S_1(T) \ge a_n,\ S_2(T) \ge b_n\right) \ge \epsilon \quad \text{for } n \in \mathbb{N}.$$

We can assume $a_n/\mu_1 \ge b_n/\mu_2$ for any $n \in \mathbb{N}$, since otherwise we can take a sub-sequence, and the following argument will still go through. Thus,

$$\epsilon \ \le\ p_n \ \le\ \mathsf{P}\left(T \le q\frac{a_n}{\mu_1}, S_1(T) \ge a_n\right),$$

which contradicts with (3.57). Thus the proof is complete. □

Remark 3.7. *Note that in* (3.58)*, there is no restriction on the way a, b approaching infinity.*

The next lemma provides an upper bound on the expectation of the first time when multiple processes

simultaneous cross given thresholds.

Lemma 3.14. *Let $L \geq 2$ and $\{\xi_\ell(n) : n \in \mathbb{N}\}_{\ell\in[L]}$ be L stochastic processes on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Define the stopping time*

$$\nu(\vec{b}) \ := \ \inf\{n \geq 1 : \ \xi_\ell(n) \geq b_\ell \ \ \textit{for every} \ \ \ell \in [L]\}$$

where $\vec{b} = \{b_1, \ldots, b_L\}$. Then for some positive $\mu_1, \ldots, \mu_L$, we have

$$\mathsf{E}[\nu(\vec{b})] \ \leq \ \max_{\ell\in[L]} \left\{\frac{b_\ell}{\mu_\ell}\right\}(1 + o(1)) \quad \textit{as} \quad \min_{\ell\in[L]}\{b_\ell\} \to \infty \tag{3.59}$$

*if **one** of the following conditions holds: (i). For each $\ell \in [L]$ and any $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} \mathsf{P}\left(\left|\frac{1}{n}\xi_\ell(n) - \mu_\ell\right| \geq \epsilon\right) < \infty.$$

(ii). For each $\ell \in [L]$, $\{\xi_\ell(n) : n \in \mathbb{N}\}$ has independent and identically distributed increment, and

$$\mathsf{P}\left(\lim_{n\to\infty} \frac{1}{n}\xi_\ell(n) = \mu_\ell\right) = 1.$$

Proof. Denote $N(\vec{b}) = \max_{\ell\in[L]}\{b_\ell/\mu_\ell\}$, and $\vec{b}_{min} = \min\{b_1, \ldots, b_L\}$.

First, assume condition (i) holds. Fix $\epsilon \in (0, 1)$, and denote $N_\epsilon(\vec{b}) = \left\lfloor N(\vec{b})/(1-\epsilon)\right\rfloor$. By definition of $\nu(\vec{b})$, we have

$$\{\nu(\vec{b}) > n\} \ \subset \ \bigcup_{\ell\in[L]} \{\xi_\ell(n) < b_\ell\}$$

By Boole's inequality, for $n > N_\epsilon(\vec{b})$,

$$\begin{aligned}
\mathsf{P}(\nu(\vec{b}) > n) &\leq \sum_{\ell\in[L]} \mathsf{P}(\xi_\ell(n) < b_\ell) \leq \sum_{\ell\in[L]} \mathsf{P}\left(\frac{1}{n}\xi_\ell(n) < \frac{b_\ell}{N_\epsilon(\vec{b}) + 1}\right) \\
&\leq \sum_{\ell\in[L]} \mathsf{P}\left(\frac{1}{n}\xi_\ell(n) < (1-\epsilon)\mu_\ell\right) \\
&\leq \sum_{\ell\in[L]} \mathsf{P}\left(\left|\frac{1}{n}\xi_\ell(n) - \mu_\ell\right| > \epsilon\mu_\ell\right),
\end{aligned}$$

where we used the fact that $n \geq N_\epsilon(\vec{b}) + 1 \geq \frac{N(\vec{b})}{1-\epsilon} \geq \frac{b_\ell}{(1-\epsilon)\mu_\ell}$. Thus

$$\begin{aligned}\mathsf{E}[\nu(\vec{b})] = \int_0^\infty \mathsf{P}(\nu(\vec{b}) > t)\, dt \;&\leq N_\epsilon(\vec{b}) + 1 + \sum_{n > N_\epsilon(\vec{b})} \mathsf{P}(\nu(\vec{b}) > n) \\ &\leq N_\epsilon(\vec{b}) + 1 + \sum_{\ell \in [L]} \sum_{n > N_\epsilon(\vec{b})} \mathsf{P}\left(\left| \frac{1}{n} \xi_\ell(n) - \mu_\ell \right| > \epsilon \mu_\ell \right)\end{aligned}$$

Due to condition (i), we have

$$\limsup_{\vec{b}_{min} \to \infty} \frac{\mathsf{E}[\nu(\vec{b})]}{N(\vec{b})} \;=\; \limsup_{\vec{b}_{min} \to \infty} (1-\epsilon) \frac{\mathsf{E}[\nu(\vec{b})]}{N_\epsilon(\vec{b})} \;\leq\; 1 - \epsilon$$

Since $\epsilon \in (0,1)$ is arbitrary, (3.59) holds.

Now assume that condition (ii) holds. Clearly, $\nu(\vec{b}) \geq \nu_\ell(b_\ell)$, where

$$\nu_\ell(b_\ell) := \inf\{n \geq 1 : \xi_\ell(n) \geq b_\ell\} \;\text{ for }\; \ell \in [L].$$

Due to condition (ii), we have

$$\liminf_{b_\ell \to \infty} \frac{\nu(\vec{b})}{b_\ell/\mu_\ell} \;\geq\; \lim_{b_\ell \to \infty} \frac{\nu_\ell(b_\ell)}{b_\ell/\mu_\ell} = 1 \;\text{ for }\; \ell \in [L],$$

which implies $\liminf_{\vec{b}_{min} \to \infty} \nu(\vec{b})/N(\vec{b}) \geq 1$. On the other hand, by the definition of $\nu(\vec{b})$, there exists $\ell' \in [L]$ such that

$$\xi_{\ell'}(\nu(\vec{b}) - 1) < b_{\ell'} \iff \frac{\xi_{\ell'}(\nu(\vec{b})) - b_{\ell'}}{\nu(\vec{b})\mu_{\ell'}} \leq \frac{\xi_{\ell'}(\nu(\vec{b})) - \xi_{\ell'}(\nu(\vec{b}) - 1)}{\nu(\vec{b})\mu_{\ell'}}.$$

Taking the minimum on the l.h.s., and maximum on the right, we have

$$\min_{\ell \in [L]} \frac{\xi_\ell(\nu(\vec{b})) - b_\ell}{\nu(\vec{b})\mu_\ell} \leq \max_{\ell \in [L]} \frac{\xi_\ell(\nu(\vec{b})) - \xi_\ell(\nu(\vec{b}) - 1)}{\nu(\vec{b})\mu_\ell}.$$

which implies

$$\frac{N(\vec{b})}{\nu(\vec{b})} \;=\; \max_{\ell \in [L]} \frac{b_\ell}{\nu(\vec{b})\mu_\ell} \;\geq\; \min_{\ell \in [L]} \frac{\xi_\ell(\nu(\vec{b}))}{\nu(\vec{b})\mu_\ell} - \max_{\ell \in [L]} \frac{\xi_\ell(\nu(\vec{b})) - \xi_\ell(\nu(\vec{b}) - 1)}{\nu(\vec{b})\mu_\ell}$$

where the last term will goes to 1 as $\vec{b}_{min} \to \infty$ due to condition (ii). Thus, $\liminf N(\vec{b})/\nu(\vec{b}) \geq 1$ as $\vec{b}_{min} \to \infty$, which together with previous reverse inequality, shows that $\nu(\vec{b})/N(\vec{b}) \to 1$ almost surely as

$\vec{b}_{min} \to \infty$. Thus, the proof would be complete if we can show the following:

$$(*) \qquad \mathcal{C}_1 = \left\{ \frac{\nu(\vec{b})}{N(\vec{b})} : \ b_1, \ldots, b_L > 0 \right\} \text{ is uniformly integrable}$$

Define $\mu_{max} = \max\{\mu_1, \ldots, \mu_L\} > 0$, $b_{max} = \max\{b_1, \ldots, b_L\}$ and

$$\nu'(c) = \inf\{n \geq 1 : \xi_\ell \geq c \ \text{ for every } \ \ell \in [L]\} \ \text{ for } \ c > 0.$$

By Theorem 3 of [24], $\mathcal{C}_2 = \{\nu'(c)/c : c > 0\}$ is uniformly integrable. Observe that

$$\nu(\vec{b}) \leq \nu'(b_{max}) \ , \ N(\vec{b}) \geq \frac{b_{max}}{\mu_{max}} \ \Rightarrow \ \frac{\nu(\vec{b})}{N(\vec{b})} \leq \mu_{max} \frac{\nu'(b_{max})}{b_{max}} \ \in \ \mu_{max} \, \mathcal{C}_2.$$

Since μ_{max} is a constant, $\mathcal{C}_1$ is dominated by a uniformly integrable family. Thus condition $(*)$ holds, and the proof is complete. □

3.14 Generalized Chernoff's lemma

In this section, we present a generalization to the Chernoff's Lemma that allows for different requirements on Type I and II errors. Consider the following simple versus simple testing problem: let $\{X_n, n \in \mathbb{N}\}$ be a sequence of independent random variables with common density f relative to some σ-finite measure ν, and for some densities f_0 and f_1,

$$\mathsf{H}_0 : f = f_0 \quad \text{vs.} \quad \mathsf{H}_1 : f = f_1.$$

Let $\mathcal{S}_n$ be the class of $\mathcal{F}_n$-measurable random variables taking value in $\{0, 1\}$, where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. For any procedure $D_n \in \mathcal{S}_n$, denote

$$p_n(D_n) := \mathsf{P}_0(D_n = 1), \quad q_n(D_n) := \mathsf{P}_1(D_n = 0),$$

where P_i is the probability measure under H_i for $i = 1, 2$. Further, denoting $Y := f_1(X_1)/f_0(X_1)$, we define

$$\Phi(z) := \sup_{\theta \in \mathbb{R}} \left\{ z\theta - \log\left(\mathsf{E}_0[Y^\theta]\right) \right\}, \ I_0 := \mathsf{E}_0[-\log(Y)], \ I_1 := \mathsf{E}_1[\log(Y)],$$

with the possibility that either I_0 or I_1 assumes ∞. We assume that there exists $h_d \in (-I_0, I_1)$ such that

$$\Phi(h_d)/d = \Phi(h_d) - h_d.$$

In particular, if $d = 1$, we can set $h_d = 0$.

Lemma 3.15. *(Generalized Chernoff's Lemma) For any $d > 0$,*

$$\lim_{n\to\infty} \inf_{D_n \in \mathcal{S}_n} \frac{1}{n} \log \left(p_n^{1/d}(D_n) + q_n(D_n) \right) = -\frac{\Phi(h_d)}{d}.$$

Remark 3.8. *When $d = 1$, since we can select $h_d = 0$, it reduces to Chernoff's Lemma [20, Corollary 3.4.6]. For $d \neq 1$, the proof is essentially the same, and we present it here for completeness.*

Proof of Lemma 3.15. Let us first fix n. Denote $\lambda(n) := \sum_{i=1}^{n} \log \frac{f_1(X_i)}{f_0(X_i)}$. By the Neyman-Pearson Lemma, it suffices to consider the tests of Neyman-Pearson form. Thus, if we denote

$$\delta_n(h) := 1 \quad \text{if and only if} \quad \frac{1}{n}\lambda(n) \geq h,$$

then we have

$$\inf_{D_n \in \mathcal{S}_n} \log \left(p_n^{1/d}(D_n) + q_n(D_n) \right) = \inf_{h \in \mathbb{R}} \log \left(p_n^{1/d}(\delta_n(h)) + q_n(\delta_n(h)) \right)$$

Since $p_n(\delta_n(h))$ is decreasing in h and $q_n(\delta_n(h))$ increasing in h, for any $h \in \mathbb{R}$, either $p_n(\delta_n(h)) \geq p_n(\delta_n(h_d))$ or $q_n(\delta_n(h)) \geq q_n(\delta_n(h_d))$. Thus

$$\inf_{D_n \in \mathcal{S}_n} \log \left(p_n^{1/d}(D_n) + q_n(D_n) \right) \geq \log \min \left\{ p_n^{1/d}(\delta_n(h_d)), q_n(\delta_n(h_d)) \right\}$$

By the Theorem 3.4.3 of [20], as $n \to \infty$,

$$\frac{1}{n} \log(p_n^{1/d}(\delta_n(h_d)) \to -\frac{\Phi(h_d)}{d}, \quad \frac{1}{n} \log(q_n(\delta_n(h_d))) \to -(\Phi(h_d) - h_d).$$

Thus by definition of h_d and sending $n \to \infty$,

$$\liminf_{n\to\infty} \inf_{D_n \in \mathcal{S}_n} \frac{1}{n} \log(p_n^{1/d}(D_n) + q_n(D_n)) \geq -\frac{\Phi(h_d)}{d}.$$

Clearly, the equality is attained by the Neyman-Pearson rule with threshold h_d, i.e., $\delta_n(h_d)$, which completes the proof. □

Chapter 4

Change acceleration and detection

4.1 Introduction

[1] The goal in the problem of quickest (or sequential) change detection (QCD) is to minimize some metric of detection delay, while controlling some metric of the false-alarm rate. In non-Bayesian formulations of this problem, the mechanism that triggers the change is considered to be completely unknown or at most partially known [48], and a worst-case analysis is adopted [42, 56]. In the Bayesian QCD, the change-point is assumed to be a random variable with given prior distribution; thus, the change mechanism in this setup is known and *exogenous* to the collected observations [48, 62, 63, 74].

In the current QCD framework, it is neither permissible nor relevant to influence the change-point. However, in certain applications the change corresponds to a desirable event that we want to not only quickly and reliably detect, but also *accelerate*. Specifically, the development of intelligent tutoring systems and e-learning environments in recent years has provided powerful instructive and assessment tools [2, 77, 78]. A major statistical problem in this context is to combine these tools efficiently in order to help a student master the skill of interest fast, and at the same time to minimize the delay in detecting mastery of the skill. Motivated by such applications, in this Chapter we propose a generalization of the *Bayesian* QCD problem whose key ingredients are (i) an experimental design aspect that influences the change-point and (ii) a minimization of the *total* expected time.

Specifically, we assume that at any given time we select a treatment (or experiment, or stimulus, depending on the application) among a number of options, and observe a response to it. Then, based on the already collected responses up to this time, we need to decide whether to stop and declare that the change has occurred, or to continue the process, in which case we have to decide the treatment for the next time-period. Therefore, in addition to a stopping rule, we also need to determine a rule for sequentially assigning treatments. We define the optimal procedure, consisting of a treatment assignment rule and a stopping rule, as the one that minimizes the average *total* number of responses subject to a constraint on the probability

[1] This chapter is based on my research posted on arXiv: Y. Song and G. Fellouris, "Change acceleration and detection", arXiv:1710.00915.

of false alarm, i.e., stopping before the change has occurred. Since the average total number of responses is (roughly) the sum of the expected time until the change happens and the expected detection delay, we refer to this problem as *change acceleration and detection*.

When there is only one treatment, i.e., without the experimental design aspect, this problem reduces to the *Bayesian* QCD problem [62, 74], where the goal is to find a stopping rule that minimizes the expected detection delay, while controlling the false alarm probability. When there are multiple treatments that not only determine the distribution of the responses before and after the change, but *also affect the change-point itself*, the treatment assignment rule plays a critical role in both accelerating and detecting the change, and the heart of the proposed problem is to resolve the trade-off between these two goals optimally.

A related problem is that of "sequential design of experiments", also known as "active hypothesis testing" or "controlled sensing" [12, 16, 49, 50]. However, the experimental design in this literature does not influence the true hypothesis, which does not change over time. Another relevant problem is the so-called "(partially observable) stochastic shortest path" problem [11, 52, 53], where the goal is to perform a series of actions in order to drive a (controlled) Markov chain to a certain absorbing state with the minimum possible cost. However, the target state in this context is assumed to be observable, i.e. the change-point is not latent, and thus there is no detection task involved.

We now state the main results of this Chapter. When the conditional probability that the change happens at some time (given that it has not happened yet) depends only on the current treatment, the proposed problem can be embedded into the framework of Markov Decision Processes (MDP) [10]. Under this simple change-point model, to which we refer as *Markovian*, we generalize the classical optimality result of Bayesian QCD [62] by showing that it is optimal to stop at the first time the posterior probability process, associated with the optimal assignment rule, exceeds a threshold (Section 4.3). However, the optimal assignment rule is obtained numerically via dynamic programming; thus, it does not provide any insights into how treatments are selected, whereas its implementation suffers from several computational issues.

Due to the restrictive modeling assumptions and computational difficulties of the MDP framework, in this Chapter we propose an intuitive scheme that is inspired by *mastery learning theory* in psychometrics [13] and is consistent with educational practice (Section 4.4). Specifically, we start with a "training" stage during which we assign a treatment that is "good" (in a sense to be specified) for accelerating the change. The training stage is stopped as soon as the posterior probability that the change has already occurred exceeds some threshold. When this happens, we switch to an "assessment" stage where we assign a treatment that is "good" (again in a sense to be specified) at detecting the change. This assessment stage is stopped as soon as either the posterior probability process exceeds a larger threshold, or a different test statistic *that*

tends to increase before the change-point exceeds a different threshold. In the former case, we terminate and declare that the change has occurred. In the latter, we switch back to a training stage and repeat the previous process until termination.

The proposed procedure has three free parameters (thresholds), for which we propose explicit values. Specifically, one of them is determined by the false alarm constraint, whereas the other two are selected in order to minimize an upper bound on the expected sample size of the proposed scheme. This upper bound applies for a general class of change-point models, beyond the Markovian case (Section 4.5). In this general framework, we show that the resulting procedure is asymptotically optimal, in the sense that it achieves the optimal expected sample size up to a first-order approximation as the false alarm probability vanishes (Section 4.6).

Therefore, the implementation and asymptotic optimality of the proposed procedure are not limited to the Markovian change-point model, as it is the case for the computation of the optimal solution using the MDP framework. We also argue that the proposed procedure is preferable for practical purposes *even in the Markovian case*. Indeed, its parameters are determined analytically, whereas the computation of the optimal procedure via dynamic programming requires extensive simulations. Moreover, a simulation study in the Markovian setup (Section 4.7) shows that its performance is very close to the optimal, suggesting that any inflicted performance loss relative to the optimal in this setup is minimal.

The structure of the remainder of the Chapter is as follows. In Section 4.2 we formulate the proposed problem. In Section 4.3 we describe a dynamic programming solution under the Markovian change-point model. In Section 4.4 we introduce the proposed scheme. In Section 4.5 we discuss an asymptotic framework that gives rise to a general class of change-point models. In Section 4.6 we show how to specify the thresholds of the proposed scheme, and establish its asymptotic optimality. We present a simulation study in Section 4.7 and conclude in Section 4.8. Omitted proofs are presented in the Section 4.9.

4.2 Problem formulation

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space hosting a discrete-time stochastic process $\{L_t, t = 0, 1 \ldots\}$. This process represents the state evolution of some system and takes values in the binary set $\{0, 1\}$ such that $L_{\Theta+t} = 1$ for every $t \geq 0$, where

$$\Theta \equiv \inf\{t \geq 0 : L_t = 1\}; \quad (\inf \emptyset = \infty).$$

That is, Θ is the time at which an irreversible change occurs, and we refer to it as the *change-point*. We assume that the process $\{L_t\}$ is latent, and thus the change-point cannot be observed. In order to infer

it, at each time $t \geq 1$ we select a treatment, X_t, and observe a response, Y_t, to it. Specifically, we assume that there is a finite number of available treatments, say K, and that each X_t is determined based on the observed responses up to time $t-1$. Thus, each X_t is a $[K]$-valued, $\mathcal{F}_{t-1}$–measurable random variable, where $[K] \equiv \{1, \ldots, K\}$ and $\mathcal{F}_t$ is the σ-algebra generated by the observed responses up to time t, i.e.,

$$\mathcal{F}_t \equiv \sigma(Y_s,\ 1 \leq s \leq t),\ t \geq 1; \quad \mathcal{F}_0 \equiv \{\emptyset, \Omega\}.$$

Our key assumption is that the unobserved change-point can be inferred by the observed responses and *influenced* by the *treatment assignment rule*, $\mathcal{X} \equiv \{X_t, t \geq 1\}$.

4.2.1 Response model and change-point model

We start with the response model. Each response is assumed to take values in some Polish space $\mathbb{Y}$ and to be conditionally independent of the past given the current state of the system and the current treatment. Specifically, for each $x \in [K]$ there are (known) densities f_x and g_x with respect to some σ-finite measure μ on $\mathcal{B}(\mathbb{Y})$ so that for every $t \geq 1$ we have

$$Y_t \mid\ X_t = x,\, L_t = i,\, \mathcal{F}_{t-1},\, \{L_s\}_{0 \leq s \leq t-1} \ \sim \begin{cases} f_x, & i = 0 \\ g_x, & i = 1. \end{cases}$$

That is, g_x (resp. f_x) is the density of a response to treatment x after (resp. before) the change. For each $x \in [K]$ we assume that the following conditions hold for the log-likelihood ratios of the response densities:

$$\begin{aligned} &\int_{\mathbb{Y}} \left(\log \frac{g_x}{f_x}\right)^2 g_x \, d\mu < \infty \ \text{ and } \ I_x \equiv \int_{\mathbb{Y}} \left(\log \frac{g_x}{f_x}\right) g_x \, d\mu > 0, \\ &\int_{\mathbb{Y}} \left(\log \frac{f_x}{g_x}\right)^2 f_x \, d\mu < \infty \ \text{ and } \ J_x \equiv \int_{\mathbb{Y}} \left(\log \frac{f_x}{g_x}\right) f_x \, d\mu > 0. \end{aligned} \tag{4.1}$$

As a result, the Kullback-Leibler divergences, I_x and J_x, between the response densities g_x and f_x are positive and finite for each $x \in [K]$.

Remark 4.1. *A common response space to all treatments is assumed without loss of generality. Indeed, if $\mathbb{Y}_x$ is the response space to treatment $x \in [K]$, then we can set $\mathbb{Y} = \mathbb{Y}_1 \times \ldots \times \mathbb{Y}_K$ and a response $y \in \mathbb{Y}_x$ to treatment x can be replaced by a new response $(y_1^*, \ldots, y_{x-1}^*, y, y_{x+1}^*, \ldots, y_K^*) \in \mathbb{Y}$, where each y_z^* is an arbitrary fixed response in $\mathbb{Y}_z$ for $z \in [K]$.*

We now turn to the change-point model. We denote by π_0 the probability that the change has occurred

before observing any response and by Π_t the *conditional* probability that the change happens at time $t \geq 1$, i.e.,

$$\pi_0 \equiv \mathsf{P}(L_0 = 1),$$
$$\Pi_t \equiv \mathsf{P}(L_t = 1 \,|\, L_{t-1} = 0, \mathcal{F}_{t-1}) = \mathsf{P}(\Theta = t \,|\, \Theta \geq t,\ \mathcal{F}_{t-1}), \quad t \geq 1.$$

We assume that Π_t depends only on the assigned treatments, $X_1, \ldots, X_t$, in the sense that there exists a function $\pi_t : [K]^t \to [0,1]$ such that

$$\Pi_t = \pi_t(X_1, \ldots, X_t), \quad t \geq 1.$$

Therefore, the change-point model is determined by the prior probability π_0 and the transition functions $\{\pi_t, t \geq 1\}$.

Remark 4.2. *The simplest change-point model arises when the transition probability at each time depends only on the current treatment, in the sense that for each $x \in [K]$ there is some constant $p_x \in [0,1]$ so that*

$$\pi_t(x_1, \ldots, x_{t-1}, x) = p_x \tag{4.2}$$

for every $(x_1, \ldots, x_{t-1}) \in [K]^{t-1}$ and $t \geq 1$. We will refer to (4.2) as the Markovian change-point model.

The postulated response and change-point models determine the evolution of the pair $\{L_t, Y_t,\, t \geq 1\}$ given the response densities $\{f_x, g_x,\, x \in [K]\}$, the transition functions $\{\pi_t,\, t \geq 0\}$, and the treatment assignment rule $\mathcal{X} = \{X_t, t \geq 1\}$.

Figure 1.2 provides a graphical illustration of the proposed model. Moreover, since $\mathbb{Y}$ is a Polish space, there exists some measurable function h and two independent sequences, $\{U_t\}$ and $\{V_t\}$, of independent, uniformly distributed in $(0,1)$ random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ such that for every $t \geq 1$ we have:

$$L_t = \mathbb{1}\{L_{t-1} = 1\} + \mathbb{1}\{L_{t-1} = 0,\ U_t \leq \Pi_t\} \ \text{ and } \ Y_t = h(X_t, L_t, V_t), \tag{4.3}$$

where $L_0 \equiv \mathbb{1}\{U_0 \leq \pi_0\}$ and $\mathbb{1}\{\cdot\}$ is the indicator function [32, Lemma 3.22].

Remark 4.3. *In this context, the change point, Θ, depends on the treatment assignment rule, $\mathcal{X}$, and we will write $\Theta_{\mathcal{X}}$ to emphasize this dependence. Similarly, we will write $\Pi_t(\mathcal{X})$ and $L_t(\mathcal{X})$ without emphasizing that Π_t and L_t depend only on the treatments assigned up to time t, $X_1, \ldots, X_t$, not the whole sequence of assigned treatments.*

4.2.2 Problem Formulation

The problem we consider is to first accelerate the change and then detect it as quickly as possible. Thus, an admissible procedure is a pair $(T, \mathcal{X})$, where $\mathcal{X} = \{X_t, t \geq 1\}$ is an adaptive *treatment assignment rule*, which determines how to assign the treatments, and T a *stopping rule*, which determines when to stop and declare that the change has occurred. Formally, T is an $\{\mathcal{F}_t\}$–stopping time, i.e., $\{T = t\} \in \mathcal{F}_t$ for every $t \geq 0$, and X_t is a $[K]$–valued, $\mathcal{F}_{t-1}$–measurable random variable for $t \geq 1$, recalling that $\{\mathcal{F}_t\}$ is the filtration generated by the observed responses.

We denote by $\mathcal{C}$ the class of all such pairs $(T, \mathcal{X})$. When T stops before the change-point $\Theta_{\mathcal{X}}$ induced by $\mathcal{X}$, a "false alarm" occurs. We are interested in procedures that control the probability of false alarm below a user-specified tolerance level $\alpha \in (0, 1)$, and denote by $\mathcal{C}_\alpha$ the corresponding class, i.e.,

$$\mathcal{C}_\alpha \equiv \{(T, \mathcal{X}) \in \mathcal{C} : \ \mathsf{P}(T < \Theta_{\mathcal{X}}) \leq \alpha \ \text{ and } \ \mathsf{P}(T < \infty, \Theta_{\mathcal{X}} < \infty) = 1\}.$$

The problem then is to find a procedure in $\mathcal{C}_\alpha$ that achieves the minimum possible expected sample size in this class,

$$\inf_{(T,\mathcal{X}) \in \mathcal{C}_\alpha} \ \mathsf{E}[T]. \tag{4.4}$$

Remark 4.4. *The expected time until stopping,* $\mathsf{E}[T]$*, can be decomposed as follows:*

$$\mathsf{E}[(T - \Theta_{\mathcal{X}})^+] \ + \ \mathsf{E}[\Theta_{\mathcal{X}}] \ - \ \mathsf{E}[(T - \Theta_{\mathcal{X}})^-]. \tag{4.5}$$

The first term is the average detection delay, which is the object of interest in the Bayesian QCD problem, the second term is the expected number of observations until the change, whereas the third one is negligible when α *is small. Therefore, minimization of the total expected sample size requires an "acceleration" of the change, in addition to a minimization of the detection delay, which is the reason why we refer to this problem as "change acceleration and detection".*

Remark 4.5. *All results in this Chapter can be established with minor modifications in the case that the problem is to minimize the sum of the first two terms in* (4.5)*.*

Remark 4.6. *When* $K = 1$*, there is no experimental design aspect, and the change-point is not affected by the observations. Thus, we recover the Bayesian QCD problem [62, 74], where the objective is to find a*

stopping rule that minimizes the average detection delay in $\mathcal{C}_\alpha$, i.e., a stopping rule in $\mathcal{C}_\alpha$ that achieves

$$\inf_{T \in \mathcal{C}_\alpha} \mathsf{E}[(T - \Theta)^+]. \tag{4.6}$$

4.2.3 Posterior odds and Shiryaev rules

We close this section by introducing some quantities and stating some related preliminary results that will be used throughout the Chapter.

For an assignment rule $\mathcal{X}$, we denote by $\Gamma_t(\mathcal{X})$ the *posterior odds* that the change has already occurred at time $t \geq 0$, i.e.,

$$\Gamma_t(\mathcal{X}) \;\equiv\; \frac{\mathsf{P}(L_t(\mathcal{X}) = 1 \,|\, \mathcal{F}_t)}{\mathsf{P}(L_t(\mathcal{X}) = 0 \,|\, \mathcal{F}_t)}, \quad t \geq 1; \quad \Gamma_0(\mathcal{X}) \equiv \frac{\pi_0}{1 - \pi_0}. \tag{4.7}$$

Moreover, we denote by $\{\widehat{\Gamma}_t(\mathcal{X}) : t \geq 0\}$ the posterior probability process that the change has already occurred, i.e.,

$$\widehat{\Gamma}_t(\mathcal{X}) \;\equiv\; \mathsf{P}(L_t(\mathcal{X}) = 1 | \mathcal{F}_t) \;=\; \frac{\Gamma_t(\mathcal{X})}{1 + \Gamma_t(\mathcal{X})}, \quad t \geq 0.$$

We denote by $T_{\mathcal{X}}(b)$ the first time the posterior odds process exceeds some fixed threshold $b > 0$, i.e.,

$$T_{\mathcal{X}}(b) = \inf\{t \geq 0 : \Gamma_t(\mathcal{X}) \geq b\}, \tag{4.8}$$

where threshold b is determined by the false alarm constraint, α. This stopping rule has been studied in the absence of experimental design ($K = 1$), where the transition functions $\{\pi_t\}$ reduce to transition probabilities.

Specifically, when the change-point has a (zero-modified) geometric distribution, i.e., there are $p, q \in (0, 1)$ so that $\pi_0 = q$ and $\pi_t = p$ for $t \geq 1$, [62] showed that $T_{\mathcal{X}}(b)$ is optimal, in the sense that it achieves (4.6) when b is chosen so that the probability of false alarm is equal to α. Further, it has been shown by [74] that $T_{\mathcal{X}}(b)$ achieves (4.6) up to a first-order asymptotic approximation as $\alpha \to 0$ when the sequence of transition probabilities, $\{\pi_t\}$, converges as $t \to \infty$ to some $p \in (0, 1)$ (in Cesàro sense).

In what follows, we refer to $T_{\mathcal{X}}$ as the *Shiryaev (stopping) rule associated with the treatment assignment rule* $\mathcal{X}$. The next Lemma shows that, for any assignment rule $\mathcal{X}$, $(\mathcal{X}, T_{\mathcal{X}}(b))$ belongs to $\mathcal{C}_\alpha$ when we set $b = (1 - \alpha)/\alpha$. Moreover, it suggests an efficient way to compute its false alarm probability via Monte Carlo simulation. We state this result in greater generality needed for the subsequent development. The proofs of the next two lemmas can be found in Section 4.9.1.

Lemma 4.1. *Let $\mathcal{X}$ be a treatment assignment rule and let S be an $\{\mathcal{F}_t\}$-stopping time such that $\mathsf{P}(S <$*

$\infty) = 1$. *Then,*

$$\mathsf{P}(S < \Theta_{\mathcal{X}} | \mathcal{F}_S) = \frac{1}{1 + \Gamma_S(\mathcal{X})}.$$

Hence, if $\mathsf{P}(\Gamma_S(\mathcal{X}) \geq b) = 1$ *for some positive* b*, then*

$$\mathsf{P}(S < \Theta_{\mathcal{X}}) \;=\; \mathsf{E}\left[1 - \widehat{\Gamma}_S(\mathcal{X})\right] \;=\; \mathsf{E}\left[\frac{1}{1 + \Gamma_S(\mathcal{X})}\right] \;\leq\; \frac{1}{1+b}.$$

The next Lemma shows that the posterior odds process admits a recursive form, an important property for both analysis and practical implementation.

Lemma 4.2. *Fix an assignment rule,* $\mathcal{X}$*. Then, for any* $t \geq 1$ *we have*

$$\Gamma_t(\mathcal{X}) \;=\; (\Gamma_{t-1}(\mathcal{X}) + \Pi_t(\mathcal{X}))\, \frac{\Lambda_t(\mathcal{X})}{1 - \Pi_t(\mathcal{X})}, \quad \textit{where} \;\; \Lambda_t(\mathcal{X}) \equiv \frac{g_{X_t}(Y_t)}{f_{X_t}(Y_t)}. \tag{4.9}$$

Hereafter, we may omit the argument $\mathcal{X}$ to lighten the notation when there is no danger of confusion.

4.3 Exact optimality in the Markovian case

In this section we obtain a procedure that is optimal, in the sense that it achieves (4.4) for any given tolerance level α, under the *Markovian* change-point model (4.2). Specifically, we generalize the optimality result in [62] by showing that the optimal stopping rule in this setup is of the form (4.8). However, the optimal assignment rule does not have an explicit form and its computation suffers from several issues. This approach is based on standard dynamic programming arguments [10], which are outlined below.

4.3.1 The main steps

Step 1. We first introduce a new objective function. Suppose that the cost is $c > 0$ for each treatment and 1 for a false alarm. We denote by π the prior belief $\mathsf{P}(L_0 = 1)$, and write P_π and E_π to emphasize this dependence. Then, the expected cost of a procedure $(T, \mathcal{X}) \in \mathcal{C}$ is

$$J_c(\pi; T, \mathcal{X}) \equiv c\, \mathsf{E}_\pi[T] + \mathsf{P}_\pi(\Theta_{\mathcal{X}} = 0) = \mathsf{E}_\pi\left[c\, T + 1 - \widehat{\Gamma}_T\right],$$

where the second equality is due to Lemma 4.1. For each $\pi \in [0, 1]$ we denote by $J_c^*(\pi)$ the infimum over all admissible procedures, i.e.,

$$J_c^*(\pi) = \inf_{(T,\mathcal{X}) \in \mathcal{C}} J_c(\pi; T, \mathcal{X}). \tag{4.10}$$

Note that the posterior probability process $\{\widehat{\Gamma}_t : t \geq 0\}$ is a sufficient statistic for the hidden process $\{L_t : t \geq 0\}$ [10], and that under (4.2), in view of recursion (4.9), we have the following recursion for posterior probability process: $\widehat{\Gamma}_0 = \pi$, and for $t \geq 1$,

$$\begin{aligned} \widehat{\Gamma}_t \;&=\; \psi(\widehat{\Gamma}_{t-1}, X_t, Y_t) \;\text{ where }\; \psi(z,x,y) \equiv \frac{(\, z + p_x(1-z)\,)g_x(y)}{\phi(y;z,x)} \\ &\text{and }\; \phi(y;z,x) \equiv (\, z + p_x(1-z)\,)g_x(y) \;+\; (1-p_x)(1-z)f_x(y). \end{aligned} \tag{4.11}$$

In addition, the conditional density of Y_t given $\mathcal{F}_{t-1}$ is $\phi(y; \widehat{\Gamma}_{t-1}, X_t)$ (see Section 4.9.2 for a proof).

Step 2. Denote by $\mathcal{J}$ the space of non-negative functions on $[0,1]$, i.e., $\mathcal{J} \equiv \{J,\ J : [0,1] \to [0,\infty]\}$, and define an operator $\mathcal{T}_c : \mathcal{J} \to \mathcal{J}$ as follows: for any $J \in \mathcal{J}$ and $z \in [0,1]$ we set

$$\mathcal{T}_c(J)(z) \equiv \min\left\{1 - z,\;\; c + \min_{x \in [K]} \int J(\psi(z,x,y))\phi(y;z,x)\mu(dy)\right\}. \tag{4.12}$$

Since the cost at each stage is positive, from standard dynamic programming theory [10, 34], it follows that the optimal cost function satisfies the Bellman equation, and can be computed by repeated application of the above operator:

$$\mathcal{T}_c(J_c^*) = J_c^*, \;\text{ and }\; \lim_{t \to \infty} \mathcal{T}_c^{\otimes\, t}(0)(z) = J_c^*(z) \;\text{ for any } z \in [0,1], \tag{4.13}$$

where 0 is the zero function in $\mathcal{J}$, and $\mathcal{T}_c^{\otimes\, t}(\cdot)$ is the operator on $\mathcal{J}$ obtained by composing $\mathcal{T}_c$ with itself for t times.

Step 3. After solving J_c^*, an optimal procedure $(T_c^*, \mathcal{X}_c^*)$, in the sense of achieving (4.10), is given by the following [10, 34]:

$$\begin{aligned} T_c^* &= \inf\{t \geq 0 : 1 - \widehat{\Gamma}_t \leq J_c^*(\widehat{\Gamma}_t)\}, \\ X_{t,c}^* &= \arg\min_{x \in [K]} \int J_c^*(\psi(\widehat{\Gamma}_{t-1}, x, y))\phi(y; \widehat{\Gamma}_{t-1}, x)\mu(dy) \;\text{ for }\; t \geq 1. \end{aligned}$$

Intuitively, $J_c^*(z)$ is the optimal "cost to go" if the current posterior probability is z. Thus, we should terminate the process the first time t that the stopping cost $1 - \widehat{\Gamma}_t$ does not exceed $J_c^*(\widehat{\Gamma}_t)$; otherwise, we should continue with the treatment that minimizes the optimal, expected future cost.

Step 4. For a given tolerance level $\alpha \in (0,1)$, if $c(\alpha)$ is selected such that

$$\mathsf{P}\left(T^*_{c(\alpha)} < \Theta_{\mathcal{X}^*_{c(\alpha)}}\right) = \alpha, \tag{4.14}$$

then the pair $(T^*_{c(\alpha)}, \mathcal{X}^*_{c(\alpha)})$ achieves (4.4), and thus is optimal for the problem of interest in this work.

The next theorem shows that the optimal stopping rule, T^*_c, is the *Shiryaev rule* associated with $\mathcal{X}^*_c$, i.e., $T_{\mathcal{X}^*_c}$ in the notation of (4.8). The proof is similar to that in the Bayesian QCD problem with a zero-modified geometric prior [62], and can be found in the Section 4.9.2.

Theorem 4.1. *For any $c > 0$ there exists constant $b_c \in [0,1]$ such that*

$$T^*_c \;=\; \inf\{\, t \geq 0 : \widehat{\Gamma}_t(\mathcal{X}^*_c) \geq b_c \,\} \;=\; \inf\{\, t \geq 0 : \Gamma_t(\mathcal{X}^*_c) \geq b_c/(1-b_c) \,\}.$$

4.3.2 Criticism

The approach described in this section can only be applied in the special case of the *Markovian* change-point model (4.2), which may be realistic for certain applications, but inappropriate for others. However, even under this particular model, this approach has several shortcomings: (i) in the repeated application of the operator $\mathcal{T}_c$, defined in (4.12), we have to discretize the state space $[0,1]$, and use interpolation to evaluate the integrand; (ii) the integral in (4.12) may be difficult to compute when the density ϕ, defined in (4.11), has a complex form; (iii) in order to find the value of $c(\alpha)$ for which the false alarm constraint (4.14) is satisfied, we need to numerically compute $(T^*_c, \mathcal{X}^*_c)$ for a wide range of values of c, and then compute *for each of them* the associated probability of false alarm via *simulation*; (iv) we do not have an explicit form for the optimal assignment rule $\mathcal{X}^*_c$, and thus there is no intuition about how treatments are selected.

This motivates us to propose in the next section a different procedure, whose design does not require any computational effort and whose performance achieves the optimal, in an asymptotic sense, but under a general framework that includes the Markovian change-point model (4.2).

4.4 A procedure inspired by mastery learning theory

4.4.1 Motivation and main idea

The proposed procedure is inspired by a pedagogical theory and approach known as *mastery learning* [13], according to which every student is able to master a skill given sufficient time and appropriate instruction.

This theory suggests training a student until there is evidence of mastery, and then assessing whether this has indeed happened. In the case of a negative assessment, the process of training/assessing is repeated until there is a positive assessment that the student has mastered the skill and is ready to move onto more advanced skills.

In this section we propose a procedure that is motivated by this idea. In order to describe it, let us assume (a bit vaguely for now, but see (4.23) for a precise definition) that treatment 1 is "good" at accelerating the change and that treatment K is "good" at detecting the change. Then, we propose starting with a *training* stage, where treatment 1 is assigned continuously in order to trigger the change as fast as possible. When we accumulate a fair amount of evidence suggesting that the change has already happened, we switch to an *assessment* stage, where treatment K is continuously assigned in order to quickly confirm or reject this hypothesis. If the data from the assessment stage suggest that the change has indeed happened, we terminate and declare that the change has occurred. Otherwise, we switch back to a training stage and the previous process is repeated until termination. We illustrate the main idea of this procedure in Fig 4.1, and continue with its formal definition.

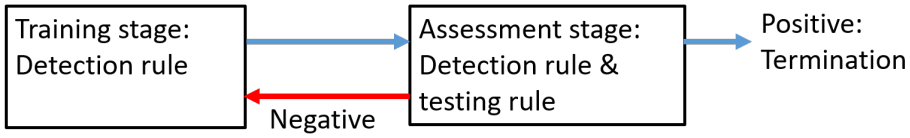


Figure 4.1: An illustration of the main idea of the proposed procedure.

4.4.2 Definition

We define a *stage* as a block of consecutive time instants at which the same treatment is assigned. We set $S_0 \equiv 0$ and for each $n \geq 1$ we denote by S_n the time that represents the *end* of the n^{th} stage, and by A_n the treatment assigned in this stage. We say that the n^{th} stage, $(S_{n-1}, S_n]$, is a *training* stage if $A_n = 1$, and an *assessment* stage if $A_n = K$.

A training stage together with its subsequent assessment stage are said to form a *cycle*, so that the m^{th} cycle is $(S_{2m-2}, S_{2m}]$, where $m \geq 1$. The proposed procedure terminates at the end of a cycle and we denote by N the number of *cycles* until stopping.

Then, the proposed procedure is defined as follows:

$$\begin{aligned}\widetilde{X}_t \equiv &\begin{cases} 1, & \text{if } t \in (S_{2m-2}, S_{2m-1}] \text{ for some } m \in \mathbb{N} \\ K, & \text{if } t \in (S_{2m-1}, S_{2m}] \text{ for some } m \in \mathbb{N} \end{cases}, \text{ for every } t \geq 1, \\ \widetilde{T} \equiv &S_{2N}. \end{aligned} \tag{4.15}$$

It remains to specify the random times $\{S_n\}$ that determine the duration of each stage, as well as the number of cycles until stopping, N. In order to do so, we need to address two questions. First, how to measure the amount of evidence supporting that the change has happened? Second, how to determine in the assessment stage that the change has *not* happened, in order to switch back to the training stage? For the first question, we introduce the following random time

$$\sigma(t; b) \equiv \inf\{s \geq 1 : \Gamma_{t+s} \geq b\}. \tag{4.16}$$

This is the number of observations required after time t by the posterior odds process (4.7), associated with the proposed assignment rule, to cross some threshold b. For the second question we introduce the random time

$$\tau(t; d) = \inf\left\{ s \geq 1 : \prod_{j=1}^{s} \frac{f_K(Y_{t+j})}{g_K(Y_{t+j})} \geq d \right\}. \tag{4.17}$$

This is a *one-sided* Sequential Probability Ratio Test (SPRT) of $L_t = 0$ against $L_t = 1$ under treatment K *if the change cannot happen in assessment stages*. However, the change may in general occur *during an assessment stage*, and this fact leads to a considerably more complicated analysis.

We now define recursively the times $\{S_n\}$ with $S_0 = 0$. Thus, at the end of the $m-1^{th}$ cycle, S_{2m-2}, we start a new training stage during which we run the change-detection procedure (4.16) with some threshold b_1 so that

$$S_{2m-1} = S_{2m-2} + \sigma(S_{2m-2}; b_1).$$

After this time, we start an assessment stage during which we run the same change-detection procedure (4.16) with some *larger* threshold $b_K > b_1$, and *at the same time* the one-sided SPRT (4.17). The assessment stage is stopped as soon as one of the two rules stops. That is,

$$S_{2m} = S_{2m-1} + \sigma(S_{2m-1}; b_K) \wedge \tau(S_{2m-1}; d),$$

where $x \wedge y = \min(x, y)$. More compactly, for each stage $n \geq 1$ we have

$$S_n = S_{n-1} + \begin{cases} \sigma(S_{n-1}; b_1), & n \text{ is odd} \\ \sigma(S_{n-1}; b_K) \wedge \tau(S_{n-1}; d), & n \text{ is even.} \end{cases} \tag{4.18}$$

Finally, we define N as the first cycle in which the the change-detection rule stops earlier than the one-sided SPRT in the assessment stage, i.e.,

$$\begin{aligned} N &\equiv \inf\{m \geq 1 : \sigma(S_{2m-1}; b_K) \leq \tau(S_{2m-1}; d)\} \\ &= \inf\{m \geq 1 : \Gamma_{S_{2m}} \geq b_K\}. \end{aligned} \tag{4.19}$$

The proposed procedure $(\widetilde{\mathcal{X}}, \widetilde{T})$ is completely determined by (4.15)–(4.19), and is illustrated graphically in Figure 4.2. In the following sections, we explain how to select the treatments in the training and assessment stages, and also how to determine thresholds b_1, b_K, d in terms of the tolerance level α.

Remark 4.7. *In view of the second equality in* (4.19)*, the proposed stopping rule, $\widetilde{T}$, resembles the Shiryaev rule with threshold b_K that is associated with $\widetilde{X}$ (recall* (4.8)*). The only difference is that the latter allows for termination at the end of a training stage, which happens if the posterior odds process at this time is not only larger than b_1, but also larger than b_K. This will be unlikely when b_K is much larger than b_1. In any case, these two stopping rules have the same asymptotic properties. We preferred to work with $\widetilde{T}$ simply because it is more intuitive and reasonable from a practical point of view to stop at the end of an assessment stage.*

4.5 The asymptotic framework

In this section we introduce a general class of change-point models for which we will be able to design the proposed scheme in the previous section, and eventually establish its asymptotic efficiency as the tolerance level $\alpha \to 0$.

Notations. $x = o(y)$ is short for $\lim(x/y) = 0$, $x = O(y)$ for $\limsup(x/y) < \infty$, $x \geq y(1 + o(1))$ for $\liminf(x/y) \geq 1$, $x \leq y(1 + o(1))$ for $\limsup(x/y) \leq 1$, and $x \sim y$ for $\lim(x/y) = 1$.

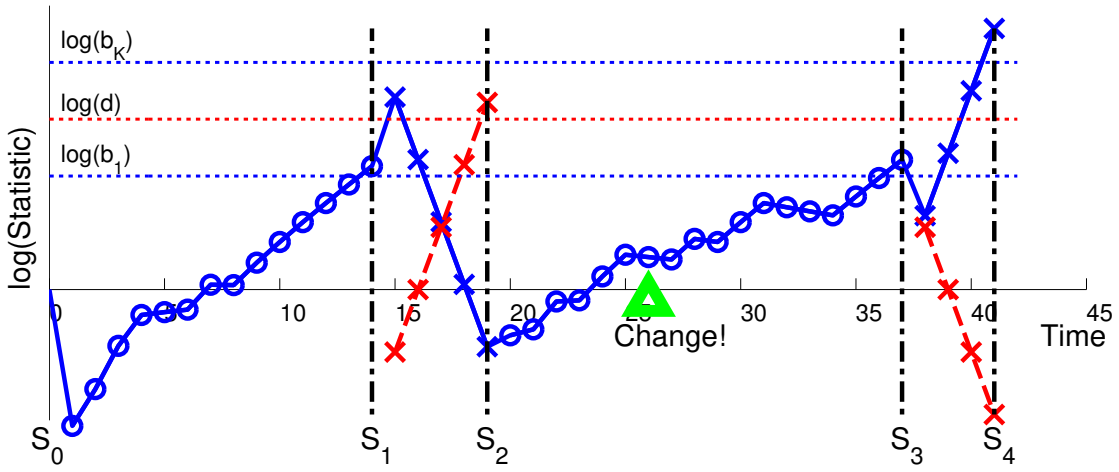


Figure 4.2: A simulation run of the proposed procedure. The circles correspond to training stages, and the crosses to assessment stages. The solid line is the logarithm of the posterior odds process, and the dashed line is the logarithm of the SPRT statistic in (4.17). In training stages, we assign treatment 1, wait until the posterior odds to cross b_1, and then switch to an assessment stage. In assessment stages, we assign treatment K, and run both the detection rule (4.16) with parameter b_K and the testing rule (4.17) with parameter d. If the testing rule stops earlier, as in the second stage of this figure, we switch back to a training stage. Otherwise, we terminate the process as in the fourth stage of this figure, where $\widetilde{T} = S_4$. Note that in this example there is no false alarm.

4.5.1 Parametrizing the transition functions by the tolerance level

Recall the decomposition (4.5) of the expected sample size $\mathsf{E}[T]$ of some pair $(T, \mathcal{X}) \in \mathcal{C}_\alpha$. Due to the false alarm constraint, the third term will be negligible as the tolerance level α goes to 0. The first term corresponds to the average detection delay and goes to infinity as $\alpha \to 0$. The second term is the expected time of change and will remain constant, thus asymptotically negligible relative to the first term, *if it is independent of* α.

Therefore, in order to conduct a more general and relevant asymptotic analysis, we need to allow the second term to go to infinity as well, maybe even faster than the first term. Thus, in what follows we *parametrize* the transition functions $\{\pi_t\}$ by α, and allow them to vanish as $\alpha \to 0$. To emphasize this parametrization, we write $\pi_t(\,\cdot\,;\alpha)$ instead of $\pi_t(\,\cdot\,)$

4.5.2 An asymptotically Markovian change-point model

In view of this enhanced asymptotic regime, we can reformulate the Markovian change-point model (4.2) as follows: for each $x \in [K]$ and $\alpha \in (0,1)$ there exists $p_x(\alpha) \in [0,1]$ such that for every $t \geq 1$ and $x_1, \ldots, x_{t-1} \in [K]$ we have

$$\pi_t(x_1, \ldots, x_{t-1}, x; \alpha) = p_x(\alpha), \tag{4.20}$$

where $p_x(\alpha)$ may go to 0 as $\alpha \to 0$. However, we will be able to analyze the proposed procedure for a more general class of change-point models, in which (4.20) is only required to hold approximately for large values of t, in the sense that

$$\sup_{\alpha>0} \sup_{z\in[K]^{t-1}} |\pi_t(z, x; \alpha) - p_x(\alpha)| \longrightarrow 0 \quad \text{as } t \to \infty. \tag{4.21}$$

This assumption is in the spirit of those imposed on the prior distribution of the change-point in the asymptotic analysis of the Bayesian QCD problem [74]. In view of the results in this literature, it is not surprising that $p_x(\alpha)$ plays a role in characterizing the detection power of treatment x.

4.5.3 Characterizing treatment quality

For each $x \in [K]$ we set

$$\mathsf{D}_x(\alpha) \equiv I_x + |\log(1 - p_x(\alpha))|,$$

where I_x is the Kullback-Leibler information number in (4.1). Moreover, for each $x \in [K]$ we denote by $\lambda_x(\alpha)$ the *expected time of the change when only treatment x is assigned*. Specifically, we denote by (x) the assignment rule under which only treatment x is assigned, i.e. $(x) \equiv \{X_t = x : t \geq 1\}$. Then,

$$\lambda_x(\alpha) \equiv \mathsf{E}[\Theta_{(x)}] = \sum_{t=0}^{\infty} \mathsf{P}(\Theta_{(x)} > t) = \sum_{t=1}^{\infty} \prod_{s=0}^{t} (1 - \pi_s(x, \ldots, x; \alpha)). \tag{4.22}$$

Without loss of generality, relabeling the treatments if necessary, we assume that

$$\lambda_1(\alpha) = \min_{x\in[K]} \lambda_x(\alpha) \quad \text{and} \quad \mathsf{D}_K(\alpha) = \max_{x\in[K]} \mathsf{D}_x(\alpha). \tag{4.23}$$

This clarifies how the treatments are selected in the proposed procedure in Section 4.4.

Remark 4.8. *In the case of the Markovian change-point model* (4.20) *we have* $\lambda_x(\alpha) = 1/p_x(\alpha)$ *and consequently* $p_1(\alpha) = \max_{x\in[K]} p_x(\alpha)$*, i.e., the treatment assigned in the training stages is the one with the highest transition probability.*

4.5.4 Additional assumptions

Finally, we need two technical assumptions. First, we assume that treatment 1 has non-trivial transition probability *whenever it is assigned*. To be more precise, let $\zeta_x(\alpha)$ denote the smallest possible transition

probability whenever treatment x is assigned, i.e.,

$$\zeta_x(\alpha) \equiv \inf_{t\geq 0} \inf_{z\in[K]^t} \pi_{t+1}(z, x; \alpha). \tag{4.24}$$

We allow $\zeta_1(\alpha)$ to vanish as $\alpha \to 0$ as long as this does not happen very fast, in the sense that

$$|\log(\zeta_1(\alpha))| \;=\; o(|\log(\alpha)|), \tag{4.25}$$

which also implies that $\zeta_1(\alpha) > 0$ for small values of α. We stress that we do *not* impose such requirement on other treatments. Thus, the transition probability may even be always 0 whenever a different treatment is assigned.

Second, we assume that all transition probabilities are bounded away from 1, which essentially implies that it is not possible to "force" the change. Specifically, let $\pi_t^*(\alpha)$ denote the maximum possible transition probability at time t, i.e.,

$$\pi_0^*(\alpha) \equiv \pi_0(\alpha), \quad \pi_t^*(\alpha) \equiv \max_{z\in[K]^t} \pi_t(z; \alpha), \quad t \geq 1. \tag{4.26}$$

Then, we assume that there is a universal constant $\delta \in (0, 1)$ such that

$$\sup_{\alpha\in(0,1)} \sup_{t\geq 0} \pi_t^*(\alpha) \leq 1 - \delta. \tag{4.27}$$

Remark 4.9. *Conditions* (4.25) *and* (4.27) *essentially exclude trivial cases. Under the Markovian change-point model* (4.20)*, they are equivalent to*

$$|\log(p_1(\alpha))| \;=\; o(|\log(\alpha)|), \tag{4.28}$$

$$\sup_{\alpha\in(0,1)} p_1(\alpha) \;<\; 1, \tag{4.29}$$

and when the transition probabilities do not depend on α, i.e., under (4.2)*, they only require that p_1 is not equal to* 0 *or* 1*.*

4.5.5 The smallest possible change-point

From (4.23) it follows that, for any given α, $\lambda_1(\alpha)$ is the smallest expected time of the change under *static assignment rules where the same treatment is always assigned.* In general, it may be possible to accelerate the change further using a non-static assignment rule. To establish a lower bound, we denote by $\Theta_*(\alpha)$ the

change-point that corresponds to the maximum transition probabilities in (4.26), i.e.,

$$\begin{aligned}&\Theta_*(\alpha) \equiv \inf\{t \geq 1 : L_t^* = 1\}, \quad \text{where} \quad L_0^* \equiv \mathbb{1}\{U_0 \leq \pi_0^*(\alpha)\} \quad \text{and} \\ &L_t^* = \mathbb{1}\{L_{t-1}^* = 1\} + \mathbb{1}\{L_{t-1}^* = 0,\, U_t \leq \pi_t^*(\alpha)\} \;\text{ for }\; t \geq 1.\end{aligned} \tag{4.30}$$

Comparing (4.30) with (4.3) we conclude that for any assignment rule $\mathcal{X}$ and tolerance level $\alpha \in (0,1)$, we have $\Theta_{\mathcal{X}}(\alpha) \geq \Theta_*(\alpha)$, and consequently $\mathsf{E}[\Theta_{\mathcal{X}}(\alpha)] \geq \lambda_*(\alpha)$, where $\lambda_*(\alpha)$ is the the expected value of $\Theta_*(\alpha)$, i.e.,

$$\lambda_*(\alpha) \equiv \mathsf{E}[\Theta_*(\alpha)] = \sum_{t=0}^{\infty} \mathsf{P}(\Theta_*(\alpha) > t) = \sum_{t=1}^{\infty} \prod_{s=0}^{t} (1 - \pi_s^*(\alpha)). \tag{4.31}$$

4.6 The main result

In this section we state and outline the proof of the main result, which is the asymptotic optimality of the proposed procedure, with an appropriate selection of thresholds, under a large class of change-point models.

First of all, from Lemma 4.1 it follows that an appropriate selection of b_K alone can guarantee the false alarm constraint. Specifically, for any given $\alpha \in (0,1)$ we have $(\widetilde{T}, \widetilde{\mathcal{X}}) \in \mathcal{C}_\alpha$ when

$$b_K = (1-\alpha)/\alpha. \tag{4.32}$$

Given this choice for b_K, the other two thresholds will be selected in order to minimize (an upper bound on) the expected sample size of the proposed scheme. Specifically, we will set

$$b_1 = \frac{1/\zeta_1(\alpha) + \log(b_K)/\mathsf{D}_K(\alpha)}{1/\mathsf{D}_1(\alpha) - 1/\mathsf{D}_K(\alpha)}, \quad d = b_1 \, \frac{1/\zeta_1(\alpha) + \log(b_K)/\mathsf{D}_K(\alpha)}{1/I_K + 1/J_K}. \tag{4.33}$$

The following theorem is the main theoretical result of this work.

Theorem 4.2. *Suppose that the response model satisfies* (4.1)*, and that the change-point model satisfies* (4.21)*,* (4.25)*,* (4.27)*.*

(i) As $\alpha \to 0$*,*

$$\inf_{(T,\mathcal{X}) \in \mathcal{C}_\alpha} \mathsf{E}[T] \;\geq\; \left(\lambda_*(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}\right)(1 + o(1)). \tag{4.34}$$

(ii) If the thresholds b_1, b_K, d *of* $(\widetilde{T}, \widetilde{\mathcal{X}})$ *are selected according to* (4.32)–(4.33)*, then* $(\widetilde{T}, \widetilde{\mathcal{X}}) \in \mathcal{C}_\alpha$ *for any given* $\alpha \in (0,1)$*, and as* $\alpha \to 0$ *we have*

$$\mathsf{E}[\widetilde{T}] \leq \left(\lambda_1(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}\right)\,(1 + o(1)). \tag{4.35}$$

Proof. We will outline the proof of (4.34) in Subsection 4.6.1 and the proof of (4.35) in Subsection 4.6.2. □

A comparison of (4.34) and (4.35) reveals that $(\widetilde{T}, \widetilde{\mathcal{X}})$ achieves the smallest possible expected sample size up to a first-order asymptotic approximation as $\alpha \to 0$ under the additional assumption that

$$\text{either (i)} \;\; \lambda_1(\alpha) \;\sim\; \lambda_*(\alpha) \qquad \text{or (ii)} \;\; \lambda_1(\alpha) = o(|\log(\alpha)|), \tag{4.36}$$

that is when the expected time of change *when only treatment 1 is assigned* is either (i) of the same order as the expectation of the smallest possible change-point, or (ii) negligible compared to the optimal expected detection delay. This is the content of the following corollary.

Corollary 4.1. *If the response model satisfies* (4.1) *and the change-point model satisfies* (4.21)*,* (4.25)*,* (4.27)*,* (4.36)*, then as* $\alpha \to 0$

$$\mathsf{E}[\widetilde{T}] \;\sim\; \lambda_1(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)} \;\sim\; \inf_{(T,\mathcal{X})\in\mathcal{C}_\alpha} \mathsf{E}[T]. \tag{4.37}$$

We now specialize our results to the case of the Markovian change-point model, using the Remark 4.9.

Corollary 4.2. *Suppose that the response model satisfies* (4.1) *and consider the Markovian change-point model* (4.20)*. Then, the asymptotic optimality property* (4.37) *holds if conditions* (4.28) *and* (4.29) *are satisfied.*

Remark 4.10. *When* (4.28) *does not hold, asymptotic optimality is achieved by the static assignment rule* (1) *and its associated Shiryaev rule,* $T_{(1)}$*.*

The following corollary states the asymptotic optimality of the proposed procedure under the original Markovian model, (4.2).

Corollary 4.3. *Suppose that the response model satisfies* (4.1) *and consider the Markovian change-point model* (4.2)*. The asymptotic optimality property* (4.37) *holds as long as the (constant) transition probability of treatment* 1*,* p_1*, is not equal to 0 or 1.*

4.6.1 Asymptotic lower bound on the optimal performance

In this subsection we establish the asymptotic lower bound (4.34) for the expected sample size of any pair $(T, \mathcal{X})$ in $\mathcal{C}_\alpha$. In view of the asymptotic framework described in Section 4.5, the change-point $\Theta_{\mathcal{X}}$ induced by $\mathcal{X}$ depends on α. However, we will simply write $\Theta_{\mathcal{X}}$ instead of $\Theta_{\mathcal{X}}(\alpha)$ to lighten the notation. Thus, for

any pair $(T, \mathcal{X})$ in $\mathcal{C}_\alpha$ we have

$$\mathsf{E}[T] \geq \mathsf{E}[T; T \geq \Theta_{\mathcal{X}}] = \mathsf{E}[\Theta_{\mathcal{X}};\ T \geq \Theta_{\mathcal{X}}] \ + \ \mathsf{E}\left[(T - \Theta_{\mathcal{X}})^+\right],$$

which implies that the infimum in (4.4) is lower bounded by

$$\inf_{(T,\mathcal{X})\in\mathcal{C}_\alpha} \mathsf{E}[\Theta_{\mathcal{X}}; T \geq \Theta_{\mathcal{X}}] \ + \ \inf_{(T,\mathcal{X})\in\mathcal{C}_\alpha} \mathsf{E}\left[(T - \Theta_{\mathcal{X}})^+\right]. \tag{4.38}$$

Therefore, it suffices to lower bound each of the two infima in (4.38). The first one represents the smallest possible average number of observations until the change when there is no false alarm. Not surprisingly, it will be lower bounded by $\lambda_*(\alpha)$, defined in (4.31), up to an asymptotically negligible term. The second one refers to the best possible *average detection delay*, which is the criterion of interest in the Bayesian QCD problem. However, existing results from this literature [74] do not apply to our setup due to the presence of an adaptive experimental design aspect. Therefore, the asymptotic lower bound for the second term in (4.38) is a novel result, for which we need to combine ideas from Bayesian QCD and sequential experimental design [16].

We now state the asymptotic lower bound for each term in (4.38).

Lemma 4.3. *(i) If* (4.25) *holds, then as* $\alpha \to 0$

$$\inf_{(T,\mathcal{X})\in\mathcal{C}_\alpha} \mathsf{E}[\Theta_{\mathcal{X}}; T \geq \Theta_{\mathcal{X}}] \geq \lambda_*(\alpha) - o(1).$$

(ii) If further (4.1), (4.21), (4.27) *hold, then as* $\alpha \to 0$

$$\inf_{(T,\mathcal{X})\in\mathcal{C}_\alpha} \mathsf{E}\left[(T - \Theta_{\mathcal{X}})^+\right] \ \geq \ \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}\,(1 + o(1)).$$

Proof. (i) Consider an arbitrary pair $(T, \mathcal{X}) \in \mathcal{C}_\alpha$. From the definition of Θ_* in (4.30) it follows that $\Theta_{\mathcal{X}} \geq \Theta_*$, and consequently

$$\mathsf{E}[\Theta_{\mathcal{X}}; T \geq \Theta_{\mathcal{X}}] \geq \ \mathsf{E}[\Theta_*; T \geq \Theta_{\mathcal{X}}] = \ \lambda_*(\alpha) - \mathsf{E}[\Theta_*; T < \Theta_{\mathcal{X}}].$$

It now remains to show that the second term in the lower bound vanishes as $\alpha \to 0$. By an application of

the Cauchy-Schwarz inequality and the definition of $\mathcal{C}_\alpha$ it follows that

$$\mathsf{E}[\Theta_*; T < \Theta_{\mathcal{X}}] \;\le\; \sqrt{\mathsf{E}\left[(\Theta_*)^2\right] \mathsf{P}(T < \Theta_{\mathcal{X}})} \;\le\; \sqrt{\mathsf{E}\left[(\Theta_*)^2\right] \alpha}.$$

By the definition of $\zeta_1(\alpha)$ in (4.24) it follows that $\Theta_{(1)}$ is stochastically dominated by a geometric random variable with parameter $\zeta_1(\alpha)$. Therefore, by assumption (4.25) we obtain

$$\mathsf{E}[(\Theta_*)^2] \le \mathsf{E}[(\Theta_{(1)})^2] \le 2/\left(\zeta_1(\alpha)\right)^2 = o(1/\alpha),$$

which completes the proof.

(ii) Fix $\epsilon, \alpha \in (0,1)$ and define

$$m_{\epsilon,\alpha} \equiv \lfloor (1-\epsilon)\,|\log(\alpha)|/\,\mathsf{D}_K(\alpha) \rfloor, \tag{4.39}$$

where $\lfloor z \rfloor$ is the largest integer that does not exceed z. For any $(T, \mathcal{X}) \in \mathcal{C}_\alpha$, by Markov's inequality we have

$$\begin{aligned}
\frac{1}{m_{\epsilon,\alpha}} \mathsf{E}\left[(T - \Theta_{\mathcal{X}})^+\right] &\ge \mathsf{P}(T \ge \Theta_{\mathcal{X}} + m_{\epsilon,\alpha}) \\
&= \mathsf{P}(T \ge \Theta_{\mathcal{X}}) - \mathsf{P}(\Theta_{\mathcal{X}} \le T < \Theta_{\mathcal{X}} + m_{\epsilon,\alpha}) \\
&\ge 1 - \alpha - \mathsf{P}(\Theta_{\mathcal{X}} \le T < \Theta_{\mathcal{X}} + m_{\epsilon,\alpha}),
\end{aligned}$$

where the last inequality follows by the definition of $\mathcal{C}_\alpha$. Therefore, it suffices to show that for any $\epsilon \in (0,1)$ we have

$$\mathsf{P}(\Theta_{\mathcal{X}} \le T \le \Theta_{\mathcal{X}} + m_{\epsilon,\alpha}) \le \delta_\epsilon(\alpha), \tag{4.40}$$

where $\delta_\epsilon(\alpha)$ does not depend on $(T, \mathcal{X})$ and vanishes as $\alpha \to 0$. Indeed, (4.40) implies

$$\inf_{(T,\mathcal{X}) \in \mathcal{C}_\alpha} \mathsf{E}\left[(T - \Theta_{\mathcal{X}})^+\right] \ge m_{\epsilon,\alpha}(1 - \alpha - \delta_\epsilon(\alpha)),$$

and the result then follows if we divide both sides by $|\log(\alpha)|/\mathsf{D}_K(\alpha)$, let $\alpha \to 0$, and then $\epsilon \to 0$.

Inequality (4.40) essentially says that, with high probability, the detection delay of a procedure in $\mathcal{C}_\alpha$ cannot be smaller than $m_{\epsilon,\alpha}$. In order to explain the idea behind the proof of this claim, let $\mathsf{R}_T^{\Theta_{\mathcal{X}}}$ denote the "likelihood ratio" statistic at time T in favor of the hypothesis that the change occurred at time $\Theta_{\mathcal{X}}$ against that the change has not happened at time T (this is defined formally in the Section). We will show that with high probability, (i) $\mathsf{R}_T^{\Theta_{\mathcal{X}}}$ cannot be smaller than (roughly) $1/\alpha$, because in this case the probability of

false alarm is not controlled below α, and (ii) $\mathsf{R}_T^{\Theta_{\mathcal{X}}}$ cannot be larger than (roughly) $1/\alpha$, because there is not sufficient time for this statistic to grow that fast if the detection delay is at most $m_{\epsilon,\alpha}$. Specifically, in Section 4.9.3 we show that for any given $\epsilon \in (0,1)$ we have

$$\mathsf{P}\left(\Theta_{\mathcal{X}} \le T < \Theta_{\mathcal{X}} + m_{\epsilon,\alpha},\ \mathsf{R}_T^{\Theta_{\mathcal{X}}} < \alpha^{-(1-\epsilon^2)}\right) \le \delta'_\epsilon(\alpha), \tag{4.41}$$

$$\mathsf{P}\left(\Theta_{\mathcal{X}} \le T < \Theta_{\mathcal{X}} + m_{\epsilon,\alpha},\ \mathsf{R}_T^{\Theta_{\mathcal{X}}} \ge \alpha^{-(1-\epsilon^2)}\right) \le \delta''_\epsilon(\alpha), \tag{4.42}$$

where $\delta'_\epsilon(\alpha)$ and $\delta''_\epsilon(\alpha)$ do not depend on $(T,\mathcal{X})$ and go to 0 as $\alpha \to 0$, which clearly implies (4.40). □

4.6.2 Upper bound on the performance of proposed procedure

We now explain why we select the thresholds b_1 and d according to (4.33) for the proposed procedure $(\widetilde{T},\widetilde{\mathcal{X}})$, defined in Section 4.4, and establish the asymptotic upper bound (4.35).

Lemma 4.4. *Suppose that* (4.1), (4.21), (4.25), (4.27) *hold. As* $\alpha \to 0$ *and* $\min\{b_1, b_K, d\} \to \infty$ *we have*

$$\mathsf{E}[\widetilde{T}] \le \mathcal{U}(b_1, b_K, d)\,(1 + o(1)),$$

where $\mathcal{U}(b_1, b_K, d)$ *is defined as follows:*

$$\left(\lambda_1(\alpha) + \frac{\log(b_K)}{\mathsf{D}_K}\right) + \left(\frac{1}{\zeta_1(\alpha)} + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)}\right)\left(\frac{1}{b_1} + \frac{1}{d}\right) + \frac{|\log(\zeta_1(\alpha))|}{\mathsf{D}_1(\alpha)} + \log(b_1)\left(\frac{1}{\mathsf{D}_1(\alpha)} - \frac{1}{\mathsf{D}_K(\alpha)}\right) + \frac{\log(d)}{b_1}\left(\frac{1}{I_K} + \frac{1}{J_K}\right).$$

Remark 4.11. *As discussed earlier, threshold* b_K *is selected according to* (4.32) *in order to guarantee the false alarm control. Given this value for* b_K*, we select* b_1 *and* d *to optimize the asymptotic upper bound* $\mathcal{U}(b_1, b_K, d)$*, which leads to the threshold values suggested in* (4.33) *(see more details in Section). With this selection of thresholds, we have*

$$\mathcal{U}(b_1, b_K, d) \ \sim\ \lambda_1(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}.$$

Outline of the proof for Lemma 4.4. We observe that

$$S_{2N} = \sum_{m=1}^{\infty} (\Delta S_{2m-1} + \Delta S_{2m})\ \mathbb{1}_{\{N \ge m\}},$$

where $\Delta S_n \equiv S_n - S_{n-1}$ is the duration of n^{th} stage, and recall that N, defined in (4.19), is the number of

cycles until stopping. Since $\{N \geq m\} \in \mathcal{F}_{S_{2m-2}} \subset \mathcal{F}_{S_{2m-1}}$, from the law of iterated expectation,

$$\mathsf{E}[\widetilde{T}] \;=\; \sum_{m=1}^{\infty} \mathsf{E}\left[\, \mathsf{E}[\Delta S_{2m-1}|\mathcal{F}_{S_{2m-2}}] + \mathsf{E}[\Delta S_{2m}|\mathcal{F}_{S_{2m-1}}];\, N \geq m\right]. \tag{4.43}$$

The first step then is to establish a non-asymptotic upper bound on the conditional expected length, $\mathsf{E}[\Delta S_n|\mathcal{F}_{S_{n-1}}]$, of each stage n, which is done in Lemma 4.6. These bounds are deterministic and do not depend on the cycle index m, which implies that the resulting upper bound for $\mathsf{E}[\widetilde{T}]$ is proportional to the expected number of cycles, $\mathsf{E}[N]$. In Lemma 4.5 we establish a non-asymptotic upper bound on $\mathsf{E}[N]$. The combination of these two bounds leads to the conclusion after letting $\alpha \to 0$. The detailed arguments and the proofs of these lemmas are presented in the Section 4.9.4. □

We start with a lemma that provides a non-asymptotic upper bound on $\mathsf{E}[N]$, which does not require any assumption on the change-point model.

Lemma 4.5. *Assume* (4.1) *holds. For any* $b_1, d > 1$, *and* $n \geq 1$,

$$\mathsf{P}(N > n) \leq \eta^n, \qquad \textit{where } \eta \equiv 1/b_1 + 1/d.$$

Consequently, $\mathsf{E}[N] \leq 1 + \eta/(1-\eta)$ *and* $\mathsf{E}[N] \to 1$ *as* $b_1 \wedge d \to \infty$.

Proof. See Section 4.9.5. □

Since $\mathsf{P}(N > 1) \leq 1/b_1 + 1/d$, this lemma implies that for large values of b_1 and d we will typically have only *one* cycle with high probability. This suggests that we need a stronger upper bound for the first training stage than the remaining ones.

Lemma 4.6. *Assume* (4.1), (4.21) *and* (4.27) *hold. For any* $\epsilon > 0$ *there exists a positive constant* C_ϵ *such that for any* $b_1, b_K, d > 0$, $\alpha \in (0,1)$, $m \in \mathbb{N}$ *we have*

$$(i) \quad \mathsf{E}\left[\Delta S_{2m-1} \mid \mathcal{F}_{S_{2m-2}}\right] \leq \left(\frac{1}{\zeta_1(\alpha)} + \frac{\log(b_1) + |\log(\zeta_1(\alpha))|}{\mathsf{D}_1(\alpha)} + C_\epsilon\right)(1+\epsilon),$$

with $1/\zeta_1(\alpha)$ *replaced by* $\lambda_1(\alpha)$ *when* $m = 1$, *and*

$$(ii) \quad \mathsf{E}[\Delta S_{2m}|\mathcal{F}_{S_{2m-1}}] \leq \left(\frac{\log(b_K/b_1)}{\mathsf{D}_K(\alpha)} + \frac{\log(d)}{b_1}\left(\frac{1}{I_K} + \frac{1}{J_K}\right) + C_\epsilon\right)(1+\epsilon).$$

Proof. See Section 4.9.6. □

Remark 4.12. *The duration of an assessment stage depends heavily on whether the change has already occurred at the end of the previous training stage. If the change has indeed happened, we would expect the change-detection rule to stop earlier than the testing rule; otherwise, we would expect the stopping to be triggered by the testing rule. This observation suggests the following decomposition for* $\mathsf{E}[\Delta S_{2m}|\mathcal{F}_{S_{2m-1}}]$,

$$\mathsf{E}\left[\Delta S_{2m}\ \mathbb{1}_{\{L_{S_{2m-1}}=1\}}|\mathcal{F}_{S_{2m-1}}\right]+\mathsf{E}\left[\Delta S_{2m}\ \mathbb{1}_{\{L_{S_{2m-1}}=0\}}|\mathcal{F}_{S_{2m-1}}\right],$$

and that we need to bound each term separately.

4.7 Simulation study

In this section we illustrate the proposed procedure and our asymptotic results in a simulation study with $K=3$ treatments under the Markovian change-point model (4.2). Specifically, we assume that the responses are Bernoulli random variables such that for every $x\in[3]$ and $t\geq 1$ we have

$$\mathsf{P}(Y_t=1|X_t=x,L_t=1)=1-f_x,\quad \mathsf{P}(Y_t=1|X_t=x,L_t=0)=f_x,$$

where $\{f_x, x\in[3]\}$ are real numbers in $(0,1)$. Moreover, we set $\pi_0=0$ and assume that the transition probability of each treatment x, p_x, does not depend on the tolerance level α. The response and transition probabilities, $\{f_x,p_x: x\in[3]\}$, are presented in Table 4.1.

We can see that treatment 1 is the best for accelerating the change (see also Remark 4.8), whereas treatment 3 is the best for detecting the change. However, while it is possible to assign exclusively treatment 1 or 2, this is not the case for treatment 3, because its transition probability is zero.

The proposed procedure (Section 4.4) uses treatment 1 in training stages and treatment 3 in assessment stages, and we will refer to it as $(1,3)$. From Corollary 4.3 it follows that this procedure is asymptotically optimal. It is also interesting to point out that using treatment 2, instead of 1, in the training stages also leads to an asymptotically optimal procedure, since the transition probability of treatment 2 is also *positive* and independent of α. We will refer to this procedure as $(2,3)$.

$x\in[3]$	f_x	p_x	D_x
1	0.45	0.1	0.125
2	0.35	0.05	0.237
3	0.25	0	0.549

Table 4.1: Response densities and transition probabilities for the three treatments.

Under the Markovian change-point model (4.2), we can also implement the optimal procedure, $(T_c^*,\mathcal{X}_c^*)$,

described in Section 4.3. Since the response space $\mathbb{Y}$ in this study is $\{0,1\}$, the integration in the operator $\mathcal{T}_c$, defined in (4.12), becomes a summation. Thus, the main challenge in the practical implementation of this approach is the computation of the constant $c(\alpha)$ for which (4.14) holds, i.e., for which the false alarm constraint is satisfied with equality. To this end, we simulate the false alarm probability of $(T_c^*, \mathcal{X}_c^*)$ for the following values of c

$$c \in \{a \cdot 10^{-b} : \ a = 1, \ldots, 9, \text{ and } \ b = 2, \ldots, 9\}.$$

Then, for any given $\alpha \in (0,1)$ we select $c(\alpha)$ to be the number in the above set with the largest error probability that does not exceed α.

Therefore, in our simulation study we compare the following procedures:

- the optimal procedure obtained via dynamic programming, $(T_c^*, \mathcal{X}_c^*)$,
- the proposed procedures, $(i, 3)$, where $i \in \{1, 2\}$, with thresholds selected according to (4.32)-(4.33),
- the procedures with a static design, (i), where $i \in \{1, 2\}$, and its associated Shiryaev stopping rule (4.8) with threshold $b = (1-\alpha)/\alpha$.

α	0.05		1E-2		1E-3		1E-5	
	Err	ESS	Err	ESS	Err	ESS	Err	ESS
optimal	0.026	21.5	9.8E-3	23.8	9.9E-4	28.3	9.6E-6	36.9
(1,3)	0.037	22.1	5.6E-3	26.9	6.9E-4	31.1	8.5E-6	39.9
(2,3)	0.027	32.8	7.0E-3	36.3	6.8E-4	41.1	6.7E-6	49.9
(1)	0.044	27.0	8.8E-3	39.9	8.8E-4	58.3	8.8E-6	95.0
(2)	0.038	32.4	7.5E-3	40.1	7.5E-4	49.9	7.4E-6	69.4

Table 4.2: Given target level α, we first determine the thresholds for each procedure, and then simulate the actual error probability (Err), and the expected sample size (ESS).

The results are summarized in Table 4.2 and Fig 4.3. In Table 4.2 we present the expected sample size (ESS) and the actual error probabilities (Err) of the above procedures for different target values of α. In Fig 4.3a we plot ESS against $-\log_{10}(\mathsf{Err})$ for each procedure, whereas in Fig 4.3b we normalize the ESS, dividing it by the associated asymptotic lower bound in (4.34), which in this context is equal to $10 - \log_{10}(\mathsf{Err})/\mathsf{D}_3$. These error probabilities were computed via the simulation method suggested in Lemma 4.1, which allowed us to set α as small as 10^{-7}.

As expected by Lemma 4.1, from Table 4.2 we observe that all procedures control the false alarm probability below the target level. For procedures (1) and (2) that employ a static design, we also observe that the ratio of the actual error probability (Err) against its target level α remains roughly constant. This finding is not surprising, as from non-linear renewal theory [76], the overshoot of a perturbed random walk

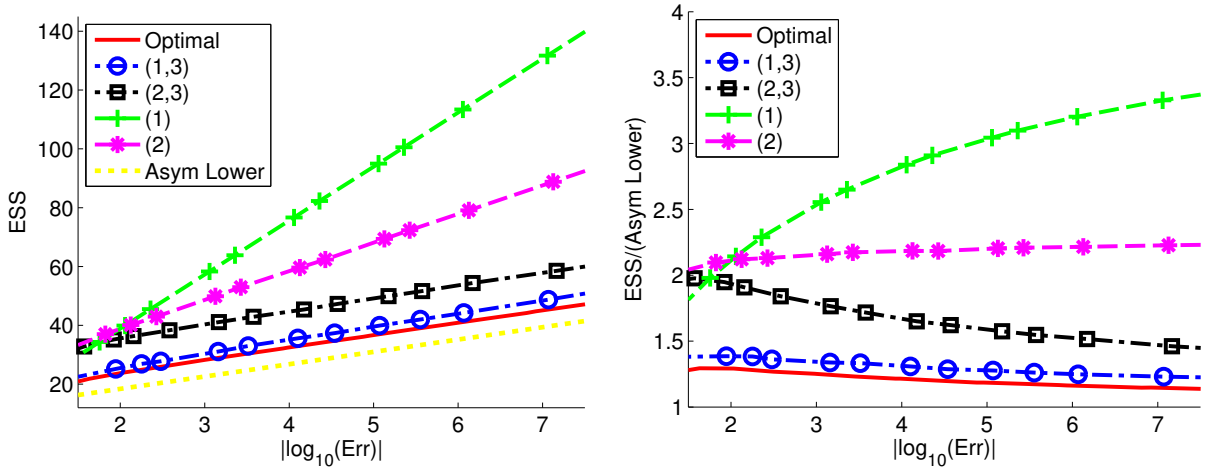


Figure 4.3: In (a), we vary the thresholds of each procedure, and plot $|\log_{10}(\text{Err})|$ vs ESS. In (b), we normalized the ESS by the asymptotic lower bound.

crossing threshold b has a limiting distribution as $b \to \infty$. On the other hand, we do not observe a similar behavior for the proposed procedure.

From Table 4.2 we also observe that the performance of the proposed procedure, $(1, 3)$, is very close to that of the optimal $(T^*(c_\alpha), \mathcal{X}^*(c_\alpha))$. Indeed, when $\alpha = 5\%$, the Err and ESS of the two procedures were roughly the same. For α equal to 1% or smaller, the Err of the optimal scheme was almost equal to α, unlike that of $(1, 3)$, and the resulting optimal ESS was consistently (roughly) 3 observations smaller than that of $(1, 3)$. Note that the performance of $(1, 3)$ in Table 4.2 was obtained by simply plugging-in the threshold values (4.32)–(4.33), whereas the implementation of the optimal scheme required extensive simulations.

The gap between the performance of $(1, 3)$ and the optimal scheme is further reduced, compared to that in Table 4.2, when both procedures are designed to have the same error probability, as depicted in Figure 4.3. Further, the gap in Figure 4.3 remains constant for small error probabilities. It suggests that the proposed procedure may enjoy an even stronger form of asymptotic optimality than the first-order property we established in this work.

From Table 4.2 and Figure 4.3a we also observe that procedure $(2, 3)$ consistently requires on average roughly 10 more samples than procedure $(1, 3)$. This is essentially the additional time required for the change under treatment 2 compared to treatment 1. As a result, the curve of (2,3) in Figure 4.3a is essentially parallel to that of (1,3), and its curve in Figure 4.3b converges to 1. On the other hand, the curves in Figure 4.3b that correspond to the "static" designs (1) and (2) do not converge to 1, which implies that these procedures fail, as expected, to be asymptotically optimal.

4.8 Conclusion

Motivated by applications in intelligent tutoring systems and e-learning environments, this Chapter proposes a generalization of the Bayesian QCD problem, where the goal is to not only detect the change as quickly as possible, but also *accelerate* it via adaptive experimental design.

Specifically, it is assumed that the sequentially collected observations are responses to treatments selected in real time. The response to each treatment has a different distribution before and after the change-point, and the change-point is influenced by the assigned treatments. The problem is to find a treatment assignment rule and a stopping rule that minimize, subject to a false alarm constraint, the expected *total* number of observations.

We obtained an exact solution to the proposed problem, via a dynamic programming approach, under the Markovian change-point model. While the optimal stopping rule admits an explicit form, this is not the case for the optimal assignment rule, whose (numerical) computation can be time-consuming and challenging. Thus in this Chapter we proposed an intuitive procedure that is easy to implement and asymptotically optimal for a large class of change-point models. Moreover, a simulation study in the Markovian case suggests that the proposed procedure is very close to the optimal.

We conclude with directions of further study: calibration of the change-point model and response models in particular applications, design and analysis of procedures that require limited information regarding the change-point and/or response models, study of the corresponding problem in the finite-horizon setup, extension to the case of multiple change-points.

4.9 Proofs

In this section, we present the omitted proofs.

4.9.1 Proofs regarding the posterior odds

Proof of Lemma 4.1. For any $t \geq 0$, by definition, we have

$$\mathsf{P}(L_t = 0 | \mathcal{F}_t) = \frac{1}{1 + \Gamma_t}.$$

Then for any $B \in \mathcal{F}_S$, we have $B \cap \{S = t\} \in \mathcal{F}_t$, and thus

$$\begin{aligned}\mathsf{P}(L_S = 0, B) &= \sum_{t=0}^{\infty} \mathsf{P}(L_t = 0, S = t, B) = \sum_{t=0}^{\infty} \mathsf{E}\left[\mathsf{P}(L_t = 0|\mathcal{F}_t); S = t, B\right] \\ &= \sum_{t=1}^{\infty} \mathsf{E}\left[\frac{1}{1+\Gamma_t}; S = t, B\right] = \mathsf{E}\left[\frac{1}{1+\Gamma_S}; B\right],\end{aligned}$$

which completes the proof by the definition of conditional expectation. □

The proof of Lemma 4.2 relies on the next Lemma, which is also crucial in establishing lower bound later. Thus, we set $\Lambda_0(\mathcal{X}) \equiv 1$, and recall the definition of $\Lambda_t(\mathcal{X})$ for $t \geq 1$ in Lemma 4.2. We denote

$$\mathsf{R}_t^s(\mathcal{X}) \equiv \Pi_s(\mathcal{X}) \prod_{j=s}^{t} \frac{\Lambda_j(\mathcal{X})}{1 - \Pi_j(\mathcal{X})}, \quad \text{for } \ t \geq s \geq 0. \tag{4.44}$$

The following lemma states that $\mathsf{R}_t^s(\mathcal{X})$ can be interpreted as the "likelihood ratio" between the hypothesis $\Theta_{\mathcal{X}} = s$ versus $\Theta_{\mathcal{X}} > t$.

Lemma 4.7. *Fix integers $t \geq s \geq 0$ and an assignment rule $\mathcal{X}$. For any non-negative measurable function $u : (\mathbb{Y}^t, \mathcal{B}(\mathbb{Y}^t)) \to [0, \infty]$, we have*

$$\mathsf{E}\left[u(Y_1, \ldots, Y_t);\ \Theta_{\mathcal{X}} = s\right] = \mathsf{E}\left[u(Y_1, \ldots, Y_t)\ \mathsf{R}_t^s(\mathcal{X});\ \Theta_{\mathcal{X}} > t\right].$$

Proof. We will only prove the case where $t \geq s \geq 1$, and other cases can be proved similarly.

Denote $y_{1:t} = (y_1, \ldots, y_t)$. Since $\mathcal{X}$ is an assignment rule, there exists a sequence of measurable function $\{x_j : j \geq 1\}$, such that $X_j = x_j(Y_{1:j})$. For any non-negative measurable function $u : \mathbb{Y}^t \to \mathbb{R}$, by an iterated conditioning argument we have

$$\begin{aligned}\mathsf{E}\left[u(Y_{1:t}); \Theta = s\right] &= \int u(y_{1:t})\, \pi_s \prod_{i=0}^{s-1}(1 - \pi_i) \prod_{i=1}^{s-1} f_{x_i}(y_i) \prod_{j=s}^{t} g_{x_j}(y_j)\ d\mu^t(y_{1:t}), \\ \mathsf{E}\left[u(Y_{1:t}); \Theta > t\right] &= \int u(y_{1:t}) \prod_{i=0}^{t}(1 - \pi_i) \prod_{i=1}^{t} f_{x_i}(y_i)\ d\mu^t(y_{1:t}),\end{aligned}$$

where we drop the arguments of $\{\pi_t\}$ and $\{x_t\}$ to simplify the notation. Since $u(\cdot)$ is arbitrary, in view of the definition (4.44) of R_t^s, we have

$$\mathsf{E}\left[u(Y_{1:t});\ \Theta = s\right] = \mathsf{E}\left[u(Y_{1:t})\, \mathsf{R}_t^s;\ \Theta > t\right].$$

which completes the proof. □

Proof of Lemma 4.2. In view of Lemma 4.7 and the definition (4.44) of R_t^s, we have for any $B \in \mathcal{F}_t$,

$$\begin{aligned}\mathsf{E}\left[L_t = 1; B\right] &= \sum_{s=0}^{t} \mathsf{P}(B, \Theta = s) = \sum_{s=0}^{t} \mathsf{E}\left[\mathsf{R}_t^s;\, B, \Theta > t\right] \\ &= \mathsf{E}\left[\sum_{s=0}^{t} \mathsf{R}_t^s;\, B, L_t = 0\right] = \mathsf{E}\left[\mathsf{P}(L_t = 0|\mathcal{F}_t) \sum_{s=0}^{t} \mathsf{R}_t^s;\, B\right].\end{aligned}$$

Thus by the definition of conditional expectation, we have

$$\mathsf{P}(L_t = 1|\mathcal{F}_t) = \mathsf{P}(L_t = 0|\mathcal{F}_t) \sum_{s=0}^{t} \mathsf{R}_t^s.$$

Thus in view of the definition (4.7) of the posterior odds, we have

$$\Gamma_t = \sum_{s=0}^{t} \mathsf{R}_t^s = \sum_{s=0}^{t} \Pi_s \prod_{j=s}^{t} \frac{\Lambda_j}{1 - \Pi_j}$$

Then simple algebra shows that the statistics $\{\Gamma_t, t \geq 0\}$ admit the recursive form (4.9). □

4.9.2 Proofs regarding the dynamic programming approach

Proof of the conditional density in (4.11). Fix some $t \geq 1$. For any $B \in \mathcal{B}(\mathbb{Y})$, we have

$$\begin{aligned}\mathsf{P}(Y_t \in B|\mathcal{F}_{t-1}) &= \mathsf{P}(Y_t \in B, L_t = 1, L_{t-1} = 1|\mathcal{F}_{t-1}) \\ &\quad + \mathsf{P}(Y_t \in B, L_t = 1, L_{t-1} = 0|\mathcal{F}_{t-1}) + \mathsf{P}(Y_t \in B, L_t = 0, L_{t-1} = 0|\mathcal{F}_{t-1}).\end{aligned}$$

Denote the three terms on the right hand side by I, II, and III. Then

$$\begin{aligned}\text{III} &= \int_B f_{X_t}(y)\mathsf{P}(L_t = 0, L_{t-1} = 0|\mathcal{F}_{t-1})\,\mu(dy) \\ &= \int_B f_{X_t}(y)(1 - p_{X_t})\mathsf{P}(L_{t-1} = 0|\mathcal{F}_{t-1})\,\mu(dy) \\ &= \int_B f_{X_t}(y)(1 - p_{X_t})(1 - \widehat{\Gamma}_{t-1})\,\mu(dy).\end{aligned}$$

By similar argument, we have

$$\text{I} = \int_B g_{X_t}(y)\widehat{\Gamma}_{t-1}\,\mu(dy), \quad \text{II} = \int_B g_{X_t}(y)p_{X_t}(1 - \widehat{\Gamma}_{t-1})\,\mu(dy).$$

Combining three terms, we have $\mathsf{P}(Y_t \in B|\mathcal{F}_{t-1}) = \int_B \phi(y; \widehat{\Gamma}_{t-1}, X_t)\mu(dy)$, which completes the proof. □

The proof of Theorem 4.1 relies on the following Lemma.

Lemma 4.8. *For any $c > 0$, J_c^* is a concave function on $[0, 1]$.*

Proof. Since point-wise limit operation preserves concavity, in view of (4.13), it suffices to show that if $J \in \mathcal{J}$ is concave, so is $\mathcal{T}_c(J)$. Since point-wise minimum and integration operations preserve concavity and $z \mapsto (1 - z)$ is a concave function, in view of the definition (4.12) of $\mathcal{T}_c$, it suffices to show that for any $x \in [K]$, $y \in \mathbb{Y}$ and concave function $J \in \mathcal{J}$, the following function is concave:

$$z \mapsto J(\psi(z, x, y))\phi(y; z, x) \quad \text{for } z \in [0, 1]. \tag{4.45}$$

With x and y fixed, to simplify notation, denote $\xi(z) \equiv (z + p_x(1 - z))g_x(y)$, and thus by (4.11), $\psi(z) = \xi(z)/\phi(z)$.

Pick any $0 \le z_1 \le z_2 \le 1$, $\gamma \in (0, 1)$. Denote $z' = \gamma z_1 + (1 - \gamma)z_2$. Then

$$\xi(z') = \gamma\xi(z_1) + (1 - \gamma)\xi(z_2), \quad \phi(z') = \gamma\phi(z_1) + (1 - \gamma)\phi(z_2).$$

By concavity of J, we have

$$\begin{aligned} & \gamma J\left(\frac{\xi(z_1)}{\phi(z_1)}\right)\phi(z_1) + (1 - \gamma)J\left(\frac{\xi(z_2)}{\phi(z_2)}\right)\phi(z_2) \\ = \ & \phi(z')\left(\frac{\gamma\phi(z_1)}{\phi(z')}J\left(\frac{\xi(z_1)}{\phi(z_1)}\right) + \frac{(1 - \gamma)\phi(z_2)}{\phi(z')}J\left(\frac{\xi(z_2)}{\phi(z_2)}\right)\right) \\ \le \ & \phi(z')J\left(\frac{\gamma\xi(z_1) + (1 - \gamma)\xi(z_2)}{\phi(z')}\right) = \phi(z')J\left(\frac{\xi(z')}{\phi(z')}\right), \end{aligned}$$

which implies the concavity of (4.45), and thus completes the proof. □

Proof of Theorem 4.1. From the definition of T_c^*, it has the following equivalent form:

$$T_c^* = \inf\{t \ge 0 : \widehat{\Gamma}_t \in B_c\}, \quad \text{where } \ B_c = \{z \in [0, 1] : J_c^*(z) - (1 - z) \ge 0\}.$$

By Lemma 4.8, J_c^* is concave, and thus so is $z \mapsto J_c^*(z) - (1 - z)$, which implies that the set B_c is convex, and thus is an interval in $[0, 1]$. Due to concavity, J_c^* is continuous, which implies that B_c is a closed interval.

Clearly, $J_c^*(1) = 0$, and thus $1 \in B_c$ and B_c is of form $[b_c, 1]$ for some $b_c \in [0, 1]$, which completes the proof. □

4.9.3 Proofs in Subsection 4.6.1

Due to the assumption (4.27) and from the definition (4.23), we have that for any $\alpha > 0$,

$$0 < I^* \leq \mathsf{D}_K(\alpha) \leq I^* + |\log(\delta)| < \infty, \quad \text{where} \quad I^* = \max_{x \in [K]} I_x. \tag{4.46}$$

Further, recall the definition of R_t^s in (4.44) and $m_{\epsilon,\alpha}$ in (4.39).

Proof of (4.41) *in Lemma 4.3.* Fix $(T, \mathcal{X}) \in \mathcal{C}_\alpha$ and write Θ instead of $\Theta_{\mathcal{X}}$ for simplicity of notation. By definition, $\mathsf{P}(T < \Theta) \leq \alpha$. Observe that

$$\begin{aligned}
\Delta \;&\equiv\; \mathsf{P}\left(\Theta \leq T < \Theta + m_{\epsilon,\alpha},\; \mathsf{R}_T^{\Theta} < \alpha^{-(1-\epsilon^2)}\right) \\
&=\; \sum_{s=0}^{\infty} \sum_{t=s}^{s+m_{\epsilon,\alpha}-1} \mathsf{P}\left(T = t,\; \mathsf{R}_t^s < \alpha^{-(1-\epsilon^2)},\; \Theta = s\right).
\end{aligned}$$

For any $t \geq s$, $\{T = t\}$ and R_t^s are both $\mathcal{F}_t$ measurable. By Lemma 4.7,

$$\begin{aligned}
\mathsf{P}\left(T = t,\; \mathsf{R}_t^s < \alpha^{-(1-\epsilon^2)},\; \Theta = s\right) &= \mathsf{E}\left[\mathsf{R}_t^s \,;\; T = t,\; \mathsf{R}_t^s < \alpha^{-(1-\epsilon^2)},\; \Theta > t\right] \\
&\leq \alpha^{-(1-\epsilon^2)}\, \mathsf{P}(T = t, \Theta > t).
\end{aligned}$$

Putting these together, we obtain

$$\begin{aligned}
\Delta \leq\; & \alpha^{-(1-\epsilon^2)} \sum_{s=0}^{\infty} \sum_{t=s}^{s+m_{\epsilon,\alpha}-1} \mathsf{P}(T = t, \Theta > t) \\
\leq\; & \alpha^{-(1-\epsilon^2)}\, m_{\epsilon,\alpha} \sum_{t=0}^{\infty} \mathsf{P}(T = t, \Theta > t) \\
= \alpha^{-(1-\epsilon^2)}\, & m_{\epsilon,\alpha}\, \mathsf{P}(T < \Theta) \leq \alpha^{\epsilon^2}\, m_{\epsilon,\alpha},
\end{aligned}$$

and the upper bound goes to 0 as $\alpha \to 0$, since due to (4.46),

$$m_{\epsilon,\alpha} \leq \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)} \leq \frac{|\log(\alpha)|}{I^*} = o(\alpha^{\epsilon^2}).$$

□

In the remainder of this section, we focus on the proof of (4.42) in Lemma 4.3. We start with a few

observations. First, we set

$$\widehat{\Lambda}_0 \equiv \log(\Lambda_0) = 0, \quad \widehat{\Lambda}_t \equiv \log(\Lambda_t) = \log\left(\frac{g_{X_t}(Y_t)}{f_{X_t}(Y_t)}\right) \quad \text{for } t \geq 1,$$

where $\{\Lambda_t : t \geq 1\}$ are defined in (4.9) and $\Lambda_0 = 1$.

Note that the treatments and the responses start from time 1, and X_0 is undefined. We further define

$$X_0 \equiv 0, \quad I_0 \equiv 0.$$

Note that $X_t \in [K]$ for any $t \geq 1$, and I_x is defined in (4.1) for $x \in [K]$.

Lemma 4.9. *Assume* (4.1) *holds. Fix any assignment rule $\mathcal{X}$, and we write Θ for $\Theta_{\mathcal{X}}$ for simplicity of notation. For any integer $t \geq 0$, we have*

$$\begin{aligned} &\mathsf{E}\left[\left(\widehat{\Lambda}_{\Theta+t} - I_{X_{\Theta+t}}\right)^2\right] \leq V^* < \infty, \quad \textit{where} \\ &V^* = \max_{x\in[K]}\{V_x\} \quad \textit{and} \quad V_x = \int_{\mathbb{Y}} \left(\log\frac{g_x}{f_x} - I_x\right)^2 g_x d\mu. \end{aligned} \tag{4.47}$$

Proof. Observe that the quantity of interest is equal to the following

$$\begin{aligned} &\sum_{s=0}^{\infty}\sum_{x=0}^{K} \mathsf{E}\left[\left(\widehat{\Lambda}_{s+t} - I_{X_{s+t}}\right)^2; \Theta = s, X_{s+t} = x\right] \\ \leq &\sum_{s=0}^{\infty}\sum_{x=0}^{K} \mathsf{P}(\Theta = s, X_{s+t} = x)\, \mathsf{E}\left[\left(\widehat{\Lambda}_{s+t} - I_x\right)^2 | \Theta = s, X_{s+t} = x\right] \\ \leq &V^* \sum_{s=0}^{\infty}\sum_{x=0}^{K} \mathsf{P}(\Theta = s, X_{s+t} = x) = V^*, \end{aligned}$$

where we used the fact that $L_{s+t} = 1$ on the event $\{\Theta = s\}$. □

Let us denote

$$\mathcal{H}_t \equiv \sigma(U_s, V_s : 0 \leq s \leq t), \quad \text{for } t \geq 0, \tag{4.48}$$

which includes all the randomness in the dynamic system (4.3) up to time t. Although $\{\mathcal{H}_t\}$ is not observable, it serves as a convenient analytic device. Clearly, $\mathcal{F}_t \subset \mathcal{H}_t$, and thus any $\{\mathcal{F}_t\}$-stopping time is $\{\mathcal{H}_t\}$-stopping time. Also, $\Theta_{\mathcal{X}}$ is an $\{\mathcal{H}_t\}$-stopping time.

Lemma 4.10. *Assume* (4.1) *holds. Fix any assignment rule $\mathcal{X}$, and we write Θ for $\Theta_{\mathcal{X}}$ for simplicity of*

notation. Then the process

$$\left\{ M_{\Theta+t} \equiv \sum_{j=\Theta}^{\Theta+t} \left(\widehat{\Lambda}_j - I_{X_j} \right) \ : \ t \geq 0 \right\} \tag{4.49}$$

is a square integrable martingale w.r.t. $\{\mathcal{H}_{\Theta+t} : t \geq 0\}$.

Proof. Adaptivity is obvious and square integrability is established in Lemma 4.9. For any $t \geq 1$, in view of (4.3) and since $L_{\Theta+t} = 1$, we have

$$Y_{\Theta+t} = h(X_{\Theta+t}, 1, V_{\Theta+t}).$$

Since $\Theta+t-1$ is an $\{\mathcal{H}_t\}$-stopping time, by Lemma 4.15, $V_{\Theta+t}$ is independent of $\mathcal{H}_{\Theta+t-1}$, and has distribution Unif$(0,1)$. Since $X_{\Theta+t} \in \mathcal{H}_{\Theta+t-1}$, we have

$$\mathsf{E}\left[\widehat{\Lambda}_{\Theta+t} - I_{X_{\Theta+t}} | \mathcal{H}_{\Theta+t-1}\right] = \int_{\mathbb{Y}} \log\left(\frac{g_{X_{\Theta+t}}}{f_{X_{\Theta+t}}}\right) g_{X_{\Theta+t}} d\,\mu - I_{X_{\Theta+t}} = 0,$$

which completes the proof. □

Next we study the behavior of above martingale.

Lemma 4.11. *Fix any assignment rule* $\mathcal{X}$*, and we write* Θ *for* $\Theta_{\mathcal{X}}$ *for simplicity of notation. Consider the process* $\{M_{\Theta+t} : t \geq 0\}$ *defined in* (4.49)*. Then, for any* $\epsilon > 0$ *we have*

$$\mathsf{P}\left(\frac{1}{m} \max_{0 \leq t < m} M_{\Theta+t} \geq \epsilon\right) \leq \frac{V^*}{\epsilon^2 m},$$

where $V^* < \infty$ *are defined in* (4.47)*.*

Proof. Observe that $z \longmapsto z^2$ is a convex function and $\{M_{\Theta+t} : t \geq 0\}$ is a square integrable $\{\mathcal{H}_{\Theta+t}\}$-martingale. Thus by Doob's inequality, we have

$$\mathsf{P}\left(\frac{1}{m} \max_{0 \leq t < m} M_{\Theta+t} \geq \epsilon\right) \leq \mathsf{P}\left(\max_{0 \leq t < m} (M_{\Theta+t})^2 \geq \epsilon^2 m^2\right) \leq \frac{\mathsf{E}[(M_{\Theta+m-1})^2]}{\epsilon^2 m^2}.$$

By properties of square-integrable martingale and the Lemma 4.9,

$$\mathsf{E}[(M_{\Theta+m-1})^2] = \sum_{s=0}^{m-1} \mathsf{E}\left[\left(\widehat{\Lambda}_{\Theta+s} - I_{X_{\Theta+s}}\right)^2\right] \leq m\, V^*,$$

which completes the proof. □

We can finally complete the proof of Lemma 4.3 by establishing (4.42).

Proof of (4.42) *in Lemma 4.3.* Pick any $(T, \mathcal{X}) \in \mathcal{C}_\alpha$ and write Θ for $\Theta_{\mathcal{X}}$. Observe that

$$\begin{aligned}
&\mathsf{P}(\Theta \le T < \Theta + m_{\epsilon,\alpha},\ \mathsf{R}_T^\Theta \ge \alpha^{-(1-\epsilon^2)}) \\
&\le \mathsf{P}\left(\max_{0\le t<m_{\epsilon,\alpha}} \log \mathsf{R}_{\Theta+t}^\Theta \ge (1-\epsilon^2)\,|\log(\alpha)|\right) \\
&\le \mathsf{P}\left(\frac{1}{m_{\epsilon,\alpha}} \max_{0\le t<m_{\epsilon,\alpha}} \log \mathsf{R}_{\Theta+t}^\Theta \ge (1+\epsilon)\mathsf{D}_K(\alpha)\right).
\end{aligned}$$

Next, by the definition of $\log \mathsf{R}_{\Theta+t}^\Theta$ in (4.44) it follows that

$$\begin{aligned}
\log \mathsf{R}_{\Theta+t}^\Theta &= -|\log \Pi_\Theta| + \sum_{j=\Theta}^{\Theta+t} \left(\widehat{\Lambda}_j - I_{X_j}\right) + \sum_{j=\Theta}^{\Theta+t} \left(|\log(1-\Pi_j)| + I_{X_j}\right) \\
&\le M_{\Theta+t} + \sum_{j=\Theta}^{\Theta+t} \left(|\log(1-\Pi_j)| + I_{X_j}\right).
\end{aligned}$$

By assumption (4.27), we have for any $j \ge 0$

$$|\log(1-\Pi_j)| + I_{X_j} \le |\log(\delta)| + I^* < \infty,$$

Due to assumptions (4.21) and (4.27), there exists some t_0 such that for any $j \ge t_0$, and $\alpha > 0$,

$$\begin{aligned}
|\log(1-\Pi_j)| + I_{X_j} &\le (1+\epsilon/2)\ (|\log(1-p_{X_j}(\alpha))| + I_{X_j}) \\
&\le (1+\epsilon/2)\ \mathsf{D}_K(\alpha).
\end{aligned}$$

Therefore, by these two observations it follows that for any $\alpha > 0$,

$$\begin{aligned}
\sum_{j=\Theta}^{\Theta+t} \left(|\log(1-\Pi_j)| + I_{X_j}\right) &\le \left(\sum_{j=0}^{t_0-1} + \sum_{j=\max\{t_0,\Theta\}}^{\Theta+t}\right) \left(|\log(1-\Pi_j)| + I_{X_j}\right) \\
&\le t_0\,(|\log(\delta)| + I^*) + t(1+\epsilon/2)\mathsf{D}_K(\alpha).
\end{aligned}$$

Note that the first term in the upper bound does not depend on α; thus, for sufficiently small α we have

$$t_0\,(I^* + |\log(\delta)|) \le (\epsilon/4)\, m_{\epsilon,\alpha}\, \mathsf{D}_K(\alpha),$$

and consequently

$$\log \mathsf{R}^{\Theta}_{\Theta+t} \leq M_{\Theta+t} + (\epsilon/4)\, m_{\epsilon,\alpha}\, \mathsf{D}_K(\alpha) + t(1+\epsilon/2)\, \mathsf{D}_K(\alpha),$$

which implies that for any $t < m_{\epsilon,\alpha}$

$$\frac{1}{m_{\epsilon,\alpha}} \log \mathsf{R}^{\Theta}_{\Theta+t} \leq \frac{1}{m_{\epsilon,\alpha}} M_{\Theta+t} + (1+3\epsilon/4)\, \mathsf{D}_K(\alpha).$$

Thus, by Lemma 4.11 it follows that there exists some constants C such that

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{m_{\epsilon,\alpha}} \max_{0\leq t<m_{\epsilon,\alpha}} \log \mathsf{R}^{\Theta}_{\Theta+t} \geq (1+\epsilon)\mathsf{D}_K(\alpha)\right) \\
&\leq \mathsf{P}\left(\frac{1}{m_{\epsilon,\alpha}} \max_{0\leq t<m_{\epsilon,\alpha}} M_{\Theta+t} \geq \frac{\epsilon \mathsf{D}_K(\alpha)}{4}\right) \\
&\leq \mathsf{P}\left(\frac{1}{m_{\epsilon,\alpha}} \max_{0\leq t<m_{\epsilon,\alpha}} M_{\Theta+t} \geq \frac{\epsilon I^*}{4}\right) \leq \frac{C}{m_{\epsilon,\alpha}},
\end{aligned}$$

which completes the proof. □

4.9.4 Proof of Lemma 4.4

We first finish the proof of Lemma 4.4 using Lemma 4.5 and 4.6.

Proof of Lemma 4.4. Fix some $\epsilon > 0$. In view of (4.43) and by Lemma 4.6, there exists some constant C_ϵ such that

$$\begin{aligned}
\frac{\mathsf{E}[\widetilde{T}]}{1+\epsilon} \;&\leq\; \lambda_1(\alpha) + \frac{1}{\zeta_1(\alpha)}(\mathsf{E}[N]-1) + \frac{\log(b_1) + |\log(\zeta_1(\alpha))|}{\mathsf{D}_1(\alpha)}\mathsf{E}[N] \\
&\quad + \left(\frac{\log(d)}{b_1}(\frac{1}{I_K}+\frac{1}{J_K}) + \frac{\log(b_K/b_1)}{\mathsf{D}_K(\alpha)}\right)\mathsf{E}[N] + C_\epsilon. \\
&=\; \lambda_1(\alpha) + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)} + \left(\frac{1}{\zeta_1(\alpha)} + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)}\right)(\mathsf{E}[N]-1) + \frac{|\log(\zeta_1(\alpha))|}{\mathsf{D}_1(\alpha)}\mathsf{E}[N] \\
&\quad + \log(b_1)\left(\frac{1}{\mathsf{D}_1(\alpha)} - \frac{1}{\mathsf{D}_K(\alpha)}\right)\mathsf{E}[N] + \frac{\log(d)}{b_1}\left(\frac{1}{I_K}+\frac{1}{J_K}\right)\mathsf{E}[N] + C_\epsilon.
\end{aligned}$$

Then by Lemma 4.5, as $\alpha \to 0$, which implies $\min\{b_1, b_K, d\} \to \infty$, we have

$$\limsup \frac{\mathsf{E}[\widetilde{T}]}{U(b_1, b_K, d)} \;\leq\; 1+\epsilon.$$

Since $\epsilon > 0$ is arbitrary, the proof of the first part is complete.

Now, plugging the thresholds (4.33) and (4.32) into $\mathcal{U}(b_1, b_K, d)$ and due to assumption (4.25), we have as $\alpha \to 0$

$$\begin{aligned}
&\lambda_1(\alpha) + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)} \;\sim\; \lambda_1(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}, \\
&\left(\frac{1}{\zeta_1(\alpha)} + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)}\right)\left(\frac{1}{b_1} + \frac{1}{d}\right) = O(1), \\
&\frac{|\log(\zeta_1(\alpha))|}{\mathsf{D}_1(\alpha)} = o(|\log(\alpha)|), \quad \log(b_1) = o(|\log(\alpha)|), \quad \log(d) = o(b_1).
\end{aligned}$$

where the third and fourth terms used assumption (4.25). Thus

$$\mathcal{U}(b_1, b_K, d) \;\sim\; \lambda_1(\alpha) + \frac{|\log(\alpha)|}{\mathsf{D}_K(\alpha)}.$$

□

Discussion of (4.33). Note that $b_K = \alpha/(1-\alpha)$ is fixed. Elementary calculus shows that for any fixed $x, y > 0$, we have

$$\arg\min_z \left\{\frac{x}{z} + y\log(z)\right\} = \frac{x}{y}. \tag{4.50}$$

Then for fixed b_1, we would choose

$$d = b_1 \, \frac{1/\zeta_1(\alpha) + \log(b_K)/\mathsf{D}_K(\alpha)}{1/I_K + 1/J_K}$$

Plugging in the above choice, and keeping the dominant terms related to b_1, we are left with

$$\left(\frac{1}{\zeta_1(\alpha)} + \frac{\log(b_K)}{\mathsf{D}_K(\alpha)}\right) \frac{1}{b_1} + \log(b_1)\left(\frac{1}{\mathsf{D}_1(\alpha)} - \frac{1}{\mathsf{D}_K(\alpha)}\right),$$

where we ignored $\log(d)/b_1$ term, since it is dominated by the first term above (as $\alpha \to 0$). Then again by (4.50), we would select b_1 as in (4.33). □

4.9.5 Proof of Lemma 4.5

We start with two observations that will be used repeatedly. By the definition (4.18) of $\{S_n\}$, the posterior odds exceeds threshold b_1 at the end of a training stage. Thus, by Lemma 4.1 we can control the conditional probability that the change has not happened at the end of a training stage. Specifically, for any $m \geq 1$,

$$\Gamma_{S_{2m-1}} \geq b_1 \quad \text{and} \quad \mathsf{P}\left(L_{S_{2m-1}} = 0 | \mathcal{F}_{S_{2m-1}}\right) \leq \frac{1}{1+b_1} \leq \frac{1}{b_1}. \tag{4.51}$$

Second, if the change has already occurred at the end of a training stage, then with high probability we terminate the process at the next assessment stage. This is formalized in the following Lemma.

Lemma 4.12. *For any integer $m \geq 1$, we have*

$$\mathsf{P}(B_{2m}, L_{S_{2m-1}} = 1 | \mathcal{F}_{S_{2m-1}}) \leq 1/d,$$

where $B_{2m} \equiv \{\tau(S_{2m-1}, d) \leq \sigma(S_{2m-1}, b_K)\}$.

Proof. Let us fix $m \geq 1$, and write S for S_{2m-1} for simplicity. Further, let us introduce the following system and its associated "stopping" rule:

$$Y'_t \equiv h(K, 1, V_{S+t}) \text{ for } t \geq 1, \quad \tau' \equiv \inf \left\{ t \geq 1 : \prod_{j=1}^{t} \frac{f_K(Y'_j)}{g_K(Y'_j)} \geq d \right\}.$$

where $\{V_t : t \geq 1\}$ appear in (4.3).

Observe that on the event $\{L_S = 1\}$, we have

$$Y_{S+t} = h(X_{S+t}, 1, V_{S+t}) = h(K, 1, V_{S+t}) = Y'_t \quad \text{for } 1 \leq t \leq S_{2m} - S.$$

Further, on the event B_{2m}, we have $S_{2m} - S = \tau(S, d)$. Thus

$$\tau(S, d) = \tau', \text{ on the event } B_{2m} \cap \{L_S = 1\}.$$

Finally, observe that

$$\begin{aligned}
\mathsf{P}(B_{2m}, L_S = 1 | \mathcal{F}_S) &= \mathsf{P}(B_{2m}, L_S = 1, \tau(S, d) < \infty | \mathcal{F}_S) \\
&= \mathsf{P}(B_{2m}, L_S = 1, \tau' < \infty | \mathcal{F}_S) \\
&\leq \mathsf{P}(L_S = 1, \tau' < \infty | \mathcal{F}_S) \\
&= \mathsf{E}[\mathsf{P}(\tau' < \infty | \mathcal{H}_S); L_S = 1 | \mathcal{F}_S],
\end{aligned}$$

where $\{\mathcal{H}_t : t \geq 0\}$ is defined in (4.48). By Lemma 4.15, $\{Y'_t, t \geq 1\}$ are independently and identically distributed (i.i.d.) with common density g_K, and are independent of $\mathcal{H}_S$. Thus, by Lemma 4.17,

$$\mathsf{P}(\tau' < \infty | \mathcal{H}_S) \leq 1/d,$$

which completes the proof. □

Remark 4.13. *In the above proof, for each fixed m, we introduced a hypothetical system $\{Y'_t : t \geq 1\}$ that is closely related to the actual responses after time S_{2m-1}, i.e. $\{Y_{S_{2m-1}+t} : t \geq 1\}$, associated with $\widetilde{\mathcal{X}}$. The advantage of the hypothetical system is that $\{Y'_t, t \geq 1\}$ is i.i.d., whereas $\{Y_{S_{2m-1}+t}, t \geq 1\}$ is not i.i.d. even on event that $\{L_{S_{2m-1}} = 1\}$, since the assigned treatments will vary in training and assessment stages.*

We are now ready to prove Lemma 4.5.

Proof of Lemma 4.5. For any integer $m \geq 1$ we have

$$\mathsf{P}(N > m) = \mathsf{P}(N > m-1, B_{2m}) = \mathsf{E}[\mathsf{P}(B_{2m}|\mathcal{F}_{S_{2m-1}}); N > m-1],$$

where B_{2m} is defined in Lemma 4.12. By (4.51) and Lemma 4.12, we have

$$\begin{aligned}\mathsf{P}(B_{2m}|\mathcal{F}_{S_{2m-1}}) &\leq \mathsf{P}(B_{2m}, L_{S_{2m-1}} = 1|\mathcal{F}_{S_{2m-1}}) + \mathsf{P}(L_{S_{2m-1}} = 0|\mathcal{F}_{S_{2m-1}}) \\ &\leq 1/d + 1/b_1 \equiv \eta.\end{aligned}$$

Then the proof is complete by telescoping argument. □

4.9.6 Proof of Lemma 4.6

In this subsection we prove Lemma 4.6, which establishes non-asymptotic upper bounds on the conditional expected length $\mathsf{E}[\Delta S_n|\mathcal{F}_{n-1}]$ of each stage n of the proposed procedure, $(\widetilde{T}, \widetilde{\mathcal{X}})$. The main idea of this proof is to introduce, for each stage, hypothetical systems that are coupled with the original system, i.e., the system $\{\Pi_t, L_t, X_t, Y_t, \Gamma_t : t \geq 1\}$ associated with the proposed assignment rule $\widetilde{\mathcal{X}}$.

Thus, for any integer $n \geq 1$, we set $x_n = 1$ if n is odd and $x_n = K$ if n is even, and define $\{\Pi^n_t, L^n_t, Y^n_t, \Gamma^n_t : t \geq 1\}$ to be a system that describes the hypothetical evolution of the transition probability, the latent state, the response, and the posterior odds of the original system after time S_{n-1} *if we only assign treatment x_n afterwards*. Specifically, if we write S for S_{n-1} for simplicity, then we define $L^n_0 \equiv L_S$, $\Gamma^n_0 \equiv \Gamma_S$ and for each

$t \geq 1$,

$$\begin{aligned}
\Pi_t^n &\equiv \pi_{S+t}(X_1, \ldots, X_S, x_n, \ldots, x_n), \\
L_t^n &\equiv \mathbb{1}\{L_{t-1}^n = 1\} + \mathbb{1}\{L_{t-1}^n = 0,\ U_t^n \leq \Pi_t^n\}, \\
Y_t^n &\equiv h(x_n, L_t^n, V_t^n), \\
\Gamma_t^n &\equiv (\Gamma_{t-1}^n + \Pi_t^n)\ \frac{g_{x_n}(Y_t^n)}{(1-\Pi_t^n) f_{x_n}(Y_t^n)},
\end{aligned} \tag{4.52}$$

where $(U_t^n, V_t^n) \equiv (U_{S+t}, V_{S+t})$ is the same "noise" that drives the original system after time S (see (4.3)). Then the evolution of the hypothetical system coincides in part with the n^{th} stage of the original system, in the sense that for any $1 \leq t \leq S_n - S_{n-1}$,

$$(\Pi_{S+t}, L_{S+t}, X_{S+t}, Y_{S+t}, \Gamma_{S+t}) = (\Pi_t^n, L_t^n, x_n, Y_t^n, \Gamma_t^n). \tag{4.53}$$

Furthermore, for each $n \geq 1$ we denote Θ^n to be the "change-point" of the above n^{th} hypothetical system, and ρ^n the required time, after the change-point, for the process $\{\Gamma_t^n : t \geq 1\}$ to cross threshold b_{x_n}. Specifically, for each $n \geq 1$,

$$\Theta^n \equiv \inf\{t \geq 1 : L_t^n = 1\}, \quad \rho^n \equiv \inf\{t \geq 0 : \Gamma_{\Theta^n + t}^n \geq b_{x_n}\}, \tag{4.54}$$

where ρ^n is well defined only on the event $\{\Theta^n < \infty\}$.

In order to upper bound the length of assessment stages, we will introduce another hypothetical system. Thus, for each *even* $n \geq 1$ we define

$$\begin{aligned}
\widehat{Y}_t^n &\equiv h(K, 0, V_t^n) \ \text{ for } t \geq 1, \\
\tau^n &\equiv \inf\left\{ t \geq 1 : \sum_{j=1}^{t} \log\left(\frac{f_K(\widehat{Y}_j^n)}{g_K(\widehat{Y}_j^n)} \right) \ \geq \log(d) \right\},
\end{aligned} \tag{4.55}$$

where $\{V_t^n : t \geq 1\}$ is the same "noise" that drives the original system after time S_{n-1} (see (4.3)). Then for any $t \leq (\Theta^n - 1) \wedge (S_n - S_{n-1})$, we have

$$\widehat{Y}_t^n = h(K, 0, V_t^n) = h(X_{S_{n-1}+t}, L_{S_{n-1}+t}, V_{S_{n-1}+t}) = Y_{S_{n-1}+t}, \tag{4.56}$$

and for any $t \leq (\Theta^n - 1)$,

$$\widehat{Y}_t^n = h(K, 0, V_t^n) = Y_t^n. \tag{4.57}$$

Note that compared to the original system, system (4.52) is simpler in that the treatments are fixed, whereas system (4.55) is even simpler in that both treatments and the latent state is fixed. The next Lemma shows that the length of each stage is bounded above by quantities of the hypothetical systems (4.52) and (4.55).

Lemma 4.13. *(i) For each $n \geq 1$, we have*

$$\Delta S_n \leq \Theta^n + \rho^n \, \mathbb{1}_{\{\Theta^n < \infty\}}.$$

(ii) If n is even, we also have

$$\Delta S_n \leq \tau^n + \rho^n \, \mathbb{1}_{\{\Theta^n \leq \tau^n < \infty\}}.$$

Proof. (i) For each $n \geq 1$, we define σ^n to be the first time the process Γ^n exceeds threshold b_{x_n}, i.e.,

$$\sigma^n \equiv \inf\{t \geq 1 : \Gamma^n_t \geq b_{x_n}\}.$$

In view of the definition of ρ^n in (4.54), we have $\sigma^n \leq \Theta^n + \rho^n \, \mathbb{1}_{\{\Theta^n < \infty\}}$, thus it suffices to show that $\Delta S_n \leq \sigma^n$. If the stopping in n^{th} stage is triggered by the detection rule, i.e. $\Gamma_{S_n} \geq b_{x_n}$, then we have $\Delta S_n = \sigma^n$ due to (4.53). Otherwise, the posterior odds of the original system does not cross b_K in the n^{th} stage, and thus again due to (4.53), we have $\Delta S_n < \sigma^n$. In any case, we have $\Delta S_n \leq \sigma^n$, and the proof is complete.

(ii) Consider some even number n. We focus on the event that $\{\tau^n < \infty\}$, since otherwise (ii) holds trivially. On the event that $\{\tau^n < \Theta^n\}$, in view of (4.55) and (4.56), the n^{th} stage of original system must have stopped by the time $S_{n-1} + \tau^n$, i.e.,

$$\Delta S_n \leq \tau^n \;\; \text{on the event } \{\tau^n < \Theta^n\}.$$

Then, together with (i) we have

$$\begin{aligned}
\Delta S_n &= \Delta S_n \mathbb{1}_{\{\tau^n < \Theta^n\}} + \Delta S_n \mathbb{1}_{\{\Theta^n \leq \tau^n\}} \\
&\leq \tau^n \mathbb{1}_{\{\tau^n < \Theta^n\}} + (\Theta^n + \rho^n) \mathbb{1}_{\{\Theta^n \leq \tau^n\}} \\
&\leq \tau^n \mathbb{1}_{\{\tau^n < \Theta^n\}} + (\tau^n + \rho^n) \mathbb{1}_{\{\Theta^n \leq \tau^n\}} = \tau^n + \rho^n \mathbb{1}_{\{\Theta^n \leq \tau^n\}},
\end{aligned}$$

which completes the proof of (ii). □

The next Lemma shows how to upper bound the stopping rule ρ^n, defined in (4.54), associated with the hypothetical system (4.52). Recall the definition (4.48) of $\{\mathcal{H}_t : t \geq 0\}$

Lemma 4.14. *Suppose that* (4.1), (4.21) *and* (4.27) *hold. Fix any* $\epsilon > 0$. *There exists some constant* $C_\epsilon > 0$ *such that the following two hold.*
(i) For any $n \geq 1$, *on the event* $\{\Theta^n < \infty\}$,

$$\rho^n \leq \inf\{t \geq 0 : Z_t^n \geq \log(b_{x_n}) - \log(\Gamma^n_{\Theta^n - 1} + \Pi^n_{\Theta^n}) + C_\epsilon\},$$

where $\{Z_t^n : t \geq 0\}$ *is a process after the change-point* Θ^n*:*

$$Z_t^n \equiv \sum_{s=\Theta^n}^{\Theta^n + t} \left[\log\left(\frac{g_{x_n}(Y_s^n)}{f_{x_n}(Y_s^n)}\right) + |\log(1 - p_{x_n}(\alpha))| - \frac{\epsilon I_{x_n}}{1+\epsilon}\right] \text{ for } t \geq 0.$$

(ii) Fix $n \geq 1$, *and set* $Z_{-1}^n = 0$. *On the event* $\{\Theta^n < \infty\}$, $\{Z_t^n - Z_{t-1}^n : t \geq 0\}$ *is a sequence of i.i.d. random variables that is independent of* $\mathcal{H}_{S_{n-1}+\Theta_n - 1}$, *that has positive first moment* $\mathsf{D}_{x_n}(\alpha) - \epsilon I_{x_n}/(1+\epsilon)$, *and that has finite second moment which only depends on the parity of* n.

Remark 4.14. *In view of (i) in the above lemma, to get a further upper bound on* ρ^n, *we have to get a lower bound on the term* $\log(\Gamma^n_{\Theta^n - 1} + \Pi^n_{\Theta^n})$, *which will be dealt with separately conditioned on different events.*

Proof. (i) From the definition (4.54) of ρ^n, it suffices to show that there exists $C_\epsilon > 0$ such that for any $n \geq 1$ and $t \geq 0$,

$$\log(\Gamma^n_{\Theta^n + t}) \geq Z_t^n + \log(\Gamma^n_{\Theta^n - 1} + \Pi^n_{\Theta^n}) - C_\epsilon. \tag{4.58}$$

By applying telescoping argument to the recursion (4.52) of $\{\Gamma_t^n : t \geq 0\}$,

$$\log \Gamma^n_{\Theta^n + t} \geq \sum_{s=\Theta^n}^{\Theta^n + t} \left(\log\left(\frac{g_{x_n}(Y_s^n)}{f_{x_n}(Y_s^n)}\right) + |\log(1 - \Pi_s^n)|\right) + \log(\Gamma^n_{\Theta^n - 1} + \Pi^n_{\Theta^n}).$$

Then, in order to prove (4.58) it suffices to show that there exists $C_\epsilon > 0$ such that for any $t \geq 0, n \geq 1$, we have

$$\sum_{s=\Theta^n}^{\Theta^n + t} |\log(1 - \Pi_s^n)| \geq \sum_{s=\Theta^n}^{\Theta^n + t} \left(|\log(1 - p_{x_n}(\alpha))| - \frac{\epsilon I_{x_n}}{1+\epsilon}\right) - C_\epsilon. \tag{4.59}$$

Now, by assumption (4.21) and (4.27), $|\log(1 - p_x(\alpha))| \leq |\log(\delta)|$ for any $x \in [K]$ and $\alpha > 0$, and there

exists some integer $s_\epsilon > 0$ such that for any $s \geq s_\epsilon$, $\alpha > 0$, and $x \in [K]$

$$\sup_{z \in [K]^{s-1}} |\log(1 - \pi_s(z, x; \alpha)) - \log(1 - p_x(\alpha))| < \frac{\epsilon I_x}{1+\epsilon}.$$

Thus, if we set $C_\epsilon = s_\epsilon |\log(\delta)|$, we have

$$\begin{aligned}\sum_{s=\Theta^n}^{\Theta^n+t} |\log(1 - p_{x_n}(\alpha))| &\leq \left(\sum_{s=0}^{s_\epsilon - 1} + \sum_{s=\max\{s_\epsilon, \Theta^n\}}^{\Theta^n+t} \right) |\log(1 - p_{x_n}(\alpha))| \\ &\leq C_\epsilon + \sum_{s=\Theta^n}^{\Theta^n+t} \left(|\log(1 - \Pi_s^n)| + \frac{\epsilon I_x}{1+\epsilon} \right),\end{aligned}$$

which clearly implies (4.59) and thus completes the proof of (i).

(ii). In view of (4.52) and by definition of Θ^n, we have for $t \geq 0$,

$$Y^n_{\Theta^n+t} = h(x_n, L^n_{\Theta^n+t}, V^n_{\Theta^n+t}) = h(x_n, 1, V^n_{\Theta^n+t}).$$

Due to Lemma 4.15, we have that

$$\{V^n_{\Theta^n+t} : t \geq 0\} \;=\; \{V_{S_{n-1}+\Theta^n+t} : t \geq 0\}$$

are independent, uniformly distributed in $(0, 1)$ random variables, that are independent of $\mathcal{H}_{S_{n-1}+\Theta^n-1}$. As a result, $\{Y^n_{\Theta^n+t} : t \geq 1\}$ is a sequence of i.i.d. random variables, that is independent of $\mathcal{H}_{S_{n-1}+\Theta^n-1}$ and that has common density g_{x_n}. Thus the proof is complete by Lemma 4.9. □

With above preparations, we can finally prove (i) and (ii) in Lemma 4.6.

Proof of Lemma 4.6(i). Consider the case (i) where n is odd and $x_n = 1$. We will only show the first claim, since the second can be proved by the same argument, and by using the definition (4.22) of $\lambda_1(\alpha)$.

By definition (4.24), we have for any $\alpha > 0$,

$$\log(\Gamma^n_{\Theta^n-1} + \Pi^n_{\Theta^n}) \geq \log(\Pi^n_{\Theta^n}) \geq \log(\zeta_1(\alpha)).$$

Thus by Lemma 4.13(i) and 4.14(i), we have $\Delta S_n \leq \Theta^n + \widetilde{\rho}^n$, where

$$\widetilde{\rho}^n \equiv \inf\{t \geq 0 : Z^n_t \geq \log(b_1) + |\log(\zeta_1(\alpha))| + C_\epsilon\}.$$

By the definition (4.24) of $\zeta_1(\alpha)$, given $\mathcal{F}_{S_{n-1}}$, Θ^n is dominated by a geometric random variable with parameter $\zeta_1(\alpha)$, and thus $\mathsf{E}[\Theta^n|\mathcal{F}_{S_{n-1}}] \le 1/\zeta_1(\alpha)$. Since $\mathcal{F}_{S_{n-1}} \subset \mathcal{H}_{S_{n-1}+\Theta^n-1}$, and due to Lemma 4.14(ii) and 4.16, there exists some constant C'_ϵ , such that for any b_1, α and odd $n \ge 1$

$$\begin{aligned}\mathsf{E}[\widetilde{\rho}^n|\mathcal{F}_{S_{n-1}}] &\le \frac{\log(b_1) + |\log(\zeta_1(\alpha))| + C'_\epsilon}{\mathsf{D}_1(\alpha) - \epsilon I_1/(1+\epsilon)} \\ &\le \frac{\log(b_1) + |\log(\zeta_1(\alpha))| + C'_\epsilon}{\mathsf{D}_1(\alpha)}\,(1+\epsilon)\end{aligned}$$

which completes the proof of (i). □

Proof of Lemma 4.6(ii). Now we consider the case (ii) where n is even and $x_n = K$. Recall Remark 4.12.

Notice that on the event $\{L_{S_{n-1}} = 1\}$, we have $\Theta^n = 1$. Further by (4.51) and the definition (4.52), on the event $\{L_{S_{n-1}} = 1\}$,

$$b_1 \le \Gamma_{S_{n-1}} = \Gamma_0^n = \Gamma^n_{\Theta^n-1} \;\Rightarrow\; \log(\Gamma^n_{\Theta^n-1} + \Pi^n_{\Theta^n}) \ge \log(b_1).$$

Thus, by Lemma 4.13(i) and 4.14(i), on the event $\{L_{S_{n-1}} = 1\}$, we have

$$\Delta S_n \le 1 + \inf\{t \ge 0 : Z_t^n \ge \log(b_K) - \log(b_1) + C_\epsilon\},$$

and then due to Lemma 4.14(ii) and 4.16, there exists some constant C'_ϵ such that for any $b_K, b_1, \alpha > 0$, and even $n \ge 1$,

$$\mathsf{E}[\Delta S_n|\mathcal{H}_{S_{n-1}}] \le \frac{\log(b_K/b_1) + C'_\epsilon}{\mathsf{D}_K(\alpha)}(1+\epsilon).$$

Since $\{L_{S_{n-1}} = 1\} \in \mathcal{H}_{S_{n-1}}$ and $\mathcal{F}_{S_{n-1}} \subset \mathcal{H}_{S_{n-1}}$, and by the law of iterated expectation, we have for any $b_K, b_1, \alpha > 0$, and even $n \ge 1$,

$$\frac{\mathsf{E}[\Delta S_n \mathbb{1}_{\{L_{S_{n-1}}=1\}}|\mathcal{F}_{S_{n-1}}]}{1+\epsilon} \le \mathsf{P}(L_{S_{n-1}} = 1|\mathcal{F}_{S_{n-1}}) \left(\frac{\log(b_K/b_1) + C'_\epsilon}{\mathsf{D}_K(\alpha)}\right). \tag{4.60}$$

Now, we focus on the event $\{L_{S_{n-1}} = 0\}$, and will apply part (ii) of Lemma 4.13. On the event $\{\Theta^n \le \tau^n\}$, by definition (4.55), we have

$$\prod_{j=1}^{\Theta^n-1} \frac{f_K(\widehat{Y}_j^n)}{g_K(\widehat{Y}_j^n)} < d,$$

thus, due to (4.52) and (4.57),

$$\Gamma^n_{\Theta^n-1} \geq \Gamma^n_0 \prod_{j=1}^{\Theta^n-1} \frac{g_K(Y^n_j)}{f_K(Y^n_j)} = \Gamma^n_0 \prod_{j=1}^{\Theta^n-1} \frac{g_K(\widehat{Y}^n_j)}{f_K(\widehat{Y}^n_j)} \geq b_1/d.$$

which implies that on the event $\{\Theta^n \leq \tau^n < \infty\}$ we have

$$\log(\Gamma^n_{\Theta^n-1} + \Pi^n_{\Theta^n}) \geq \log(b_1/d).$$

Then, due to Lemma 4.13(ii) and 4.14(i) we have

$$\Delta S_n \leq \tau^n + \widehat{\rho}^n \mathbb{1}_{\{\Theta^n \leq \tau^n < \infty\}},$$

where $\widehat{\rho}^n \equiv \inf\{t \geq 0 : Z^n_t \geq \log(b_K) - \log(b_1/d) + C_\epsilon\}$.

Due to Lemma 4.15, $\{\widehat{Y}^n_t : t \geq 1\}$ is a sequence of i.i.d. random variables with common density f_K, that is independent of $\mathcal{H}_{S_{n-1}}$. Further, recall the discussion on $\{Z^n_t : t \geq 0\}$ in Lemma 4.14(ii). Then by Lemma 4.16 and the law of iterated expectation, there exists some C'_ϵ such that for any even $n \geq 2$, and $\alpha > 0$,

$$\begin{aligned}
&\frac{\mathsf{E}[\tau^n \mathbb{1}_{\{L_{S_{n-1}}=0\}} | \mathcal{F}_{S_{n-1}}]}{1+\epsilon} \leq \mathsf{P}(L_{S_{n-1}} = 0 | \mathcal{F}_{S_{n-1}}) \frac{\log(d) + C'_\epsilon}{J_K}, \\
&\frac{\mathsf{E}[\widehat{\rho}^n \mathbb{1}_{\{\Theta^n \leq \tau^n,\, L_{S_{n-1}}=0\}} | \mathcal{F}_{S_{n-1}}]}{1+\epsilon} \leq \mathsf{P}(L_{S_{n-1}} = 0 | \mathcal{F}_{S_{n-1}}) \frac{\log(b_K/b_1) + \log(d) + C'_\epsilon}{\mathsf{D}_K(\alpha)}.
\end{aligned}$$

which implies (increasing C'_ϵ if necessary) that

$$\begin{aligned}
&\frac{\mathsf{E}[\Delta S_n \mathbb{1}_{\{L_{S_{n-1}}=0\}} | \mathcal{F}_{S_{n-1}}]}{1+\epsilon} \\
&\leq \mathsf{P}(L_{S_{n-1}} = 0 | \mathcal{F}_{S_{n-1}}) \frac{\log(b_K/b_1)}{\mathsf{D}_K(\alpha)} + \frac{\log(d)}{b_1}\left(\frac{1}{I_K} + \frac{1}{J_K}\right) + C'_\epsilon.
\end{aligned} \tag{4.61}$$

Finally, combining (4.60) and (4.61), we finish the proof of (ii) in Lemma 4.6. □

4.9.7 Additional lemmas

The following lemma is widely known and its proof can be found, e.g., in Theorem 4.1.3 of [23].

Lemma 4.15. *Let $\{W_t, t \geq 0\}$ be a sequence of independently and identically distributed R^d-valued random variables (d being an integer), and denote $\{\mathcal{G}_t = \sigma(W_s : 0 \leq s \leq t), t \geq 0\}$ its natural filtration. Let S be an $\{\mathcal{G}_t\}$-stopping time such that $\mathsf{P}(S < \infty) = 1$. Then $\{W_{S+t}, t \geq 1\}$ is independent of $\mathcal{G}_S$, and has the same*

distribution as $\{W_t, t \geq 0\}$.

The following result is non-asymptotic, and is due to [41].

Lemma 4.16. *Let* $\{Z_t, t \geq 1\}$ *be independently and identically distributed random variables, and* $\{S_t \equiv \sum_{s=1}^{t} Z_s, t \geq 1\}$ *the associated random walk. Denote* $T(b)$ *the first time that* $\{S_t\}$ *crosses some threshold* b*, i.e.*

$$T(b) = \inf\{t \geq 1 : S_t > b\}.$$

Assume that $\mathsf{E}[(Z_1^+)^2] < \infty$ *and* $\mathsf{E}[Z_1] > 0$*. Then for any* $b > 0$*, we have*

$$\mathsf{E}[T(b)] \;\leq\; \frac{b + \mathsf{E}[(Z_1^+)^2]/\mathsf{E}[Z_1]}{\mathsf{E}[Z_1]}.$$

The following lemma regarding the "one-sided" sequential probability ratio test follows directly from Wald's likelihood ratio identity [75].

Lemma 4.17. *Let* f *and* g *be two densities on measurable space* $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ *relative to some measure* μ*, and* $\{Y_t, t \geq 1\}$ *be a sequence of independent random variables with common density* g*. Further, define for any* $d > 0$*,*

$$\tau(d) \equiv \inf\left\{t \geq 1 : \prod_{s=1}^{t} \frac{f(Y_s)}{g(Y_s)} \geq d\right\}.$$

Then $\mathsf{P}(\tau(d) < \infty) \leq 1/d$.

References

[1] Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(1):137–144.

[2] Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer.

[3] Bartroff, J. (2017). Multiple Hypothesis Tests Controlling Generalized Error Rates for Sequential Data. *Statistica Sinica, in press.*

[4] Bartroff, J. and Lai, T. L. (2008). Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Sequential Analysis*, 27(3):254–276.

[5] Bartroff, J. and Lai, T. L. (2010). Multistage tests of multiple hypotheses. *Communications in Statistics–Theory and Methods*, 39(8-9):1597–1607.

[6] Bartroff, J. and Song, J. (2013). Sequential Tests of Multiple Hypotheses Controlling False Discovery and Nondiscovery Rates. *arXiv:1311.3350 [stat.ME].*

[7] Bartroff, J. and Song, J. (2014). Sequential tests of multiple hypotheses controlling type i and ii familywise error rates. *Journal of statistical planning and inference*, 153:100–114.

[8] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

[9] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

[10] Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA.

[11] Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595.

[12] Bessler, S. A. (1960). Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments, part i theory. Technical Report 55, Department of Statistics, Stanford University.

[13] Bloom, B. S. (1968). Learning for mastery. instruction and curriculum. regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1. *Evaluation comment*, 1(2):n2.

[14] Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, pages 1551–1579.

[15] Chan, H. P. and Lai, T. L. (2000). Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. *Annals of statistics*, pages 1638–1669.

[16] Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Statist.*, 30(3):755–770.

[17] De, S. K. and Baron, M. (2012a). Sequential bonferroni methods for multiple hypothesis testing with strong control of family-wise error rates i and ii. *Sequential Analysis*, 31(2):238–262.

[18] De, S. K. and Baron, M. (2012b). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *Journal of Statistical Planning and Inference*, 142(7):2059–2070.

[19] De, S. K. and Baron, M. (2015). Sequential tests controlling generalized familywise error rates. *Statistical Methodology*, 23:88 – 102.

[20] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Applications of mathematics. Springer.

[21] Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests. i. asymptotic optimality. *Information Theory, IEEE Transactions on*, 45(7):2448–2461.

[22] Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V. V. (2000). Multihypothesis sequential probability ratio tests. ii. accurate asymptotic expansions for the expected sample size. *Information Theory, IEEE Transactions on*, 46(4):1366–1383.

[23] Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

[24] Farrell, R. H. (1964). Limit theorems for stopped random walks. *Ann. Math. Statist.*, 35(3):1332–1343.

[25] Fellouris, G. and Tartakovsky, A. (2017). Multichannel sequential detection—part i: Non-iid data. *IEEE Transactions on Information Theory*.

[26] Foresti, G. L., Regazzoni, C. S., and Varshney, P. K. (2003). *Multisensor surveillance systems: the fusion perspective*. Springer Science & Business Media.

[27] Guo, W., He, L., Sarkar, S. K., et al. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics*, 42(3):1070–1101.

[28] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

[29] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386.

[30] Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 154–161. Springer.

[31] Hsu, P.-L. and Robbins, H. (1947). Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences*, 33(2):25–31.

[32] Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer-Verlag New York, 2 edition.

[33] Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

[34] Kumar, P. (1985). A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380.

[35] Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, pages 303–351.

[36] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.*, 33(3):1138–1154.

[37] Lehmann, E. L., Romano, J. P., and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.*, 33(3):1084–1108.

[38] Li, X., Liu, J., and Ying, Z. (2014a). Generalized sequential probability ratio test for separate families of hypotheses. *Sequential analysis*, 33(4):539–563.

[39] Li, Y., Nitinawarat, S., and Veeravalli, V. V. (2014b). Universal outlier hypothesis testing. *IEEE Transactions on Information Theory*, 60(7):4066–4082.

[40] Li, Y., Nitinawarat, S., and Veeravalli, V. V. (2014c). Universal sequential outlier hypothesis testing. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3205–3209. IEEE.

[41] Lorden, G. (1970). On excess over the boundary. *Ann. Math. Statist.*, 41(2):520–527.

[42] Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908.

[43] Lorden, G. (1973). Open-ended tests for koopman-darmois families. *Ann. Statist.*, 1(4):633–643.

[44] Lorden, G. (1977). Nearly-optimal sequential tests for finitely many parameter values. *Ann. Statist.*, pages 1–21.

[45] Malloy, M. L. and Nowak, R. D. (2014). Sequential testing for sparse recovery. *Information Theory, IEEE Transactions on*, 60(12):7862–7873.

[46] Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

[47] Mei, Y. (2008). Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks. *Information Theory, IEEE Transactions on*, 54(5):2072–2089.

[48] Moustakides, G. V. (2008). Sequential change detection revisited. *Ann. Statist.*, 36(2):787–807.

[49] Naghshvar, M. and Javidi, T. (2013). Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738.

[50] Nitinawarat, S., Atia, G. K., and Veeravalli, V. V. (2013). Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464.

[51] Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

[52] Patek, S. D. (2001). On partially observed stochastic shortest path problems. In *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)*, volume 5, pages 5050–5055 vol.5.

[53] Patek, S. D. (2007). Partially observed stochastic shortest path problems with approximate solution by neurodynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(5):710–720.

[54] Pavlov, I. (1991). Sequential procedure of testing composite hypotheses with applications to the kiefer–weiss problem. *Theory of Probability & Its Applications*, 35(2):280–292.

[55] Peña, E. A., Habiger, J. D., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.*, 39(1):556–583.

[56] Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.*, 13(1):206–227.

[57] Rappaport, T. S. et al. (1996). *Wireless communications: principles and practice*, volume 2. prentice hall PTR New Jersey.

[58] Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *Ann. Statist.*, 2(3):415–436.

[59] Romano, J. P. and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, pages 1850–1873.

[60] Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, pages 1378–1408.

[61] Shewhart, W. A. (1931). *Economic control of manufactured product*. van Nostrand.

[62] Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46.

[63] Shiryaev, A. N. (2007). *Optimal stopping rules*, volume 8. Springer Science & Business Media.

[64] Sobel, M. and Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Statist.*, 20(4):502–522.

[65] Song, Y. and Fellouris, G. (2016). Logarithmically efficient simulation for misclassification probabilities in sequential multiple testing. In *Winter Simulation Conference (WSC), 2016*, pages 314–325. IEEE.

[66] Song, Y. and Fellouris, G. (2017). Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals. *Electronic Journal of Statistics*, 11(1):338–363.

[67] Song, Y. and Fellouris, G. (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *Ann. Statist.*, 47(3):1776–1803.

[68] Stoica, G. (2007). Baum–katz–nagaev type results for martingales. *Journal of Mathematical Analysis and Applications*, 336(2):1489–1492.

[69] Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368.

[70] Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424.

[71] Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press.

[72] Tartakovsky, A. G. (1998). Asymptotic optimality of certain multihypothesis sequential tests: Non-iid case. *Statistical Inference for Stochastic Processes*, 1(3):265–295.

[73] Tartakovsky, A. G., Li, X. R., and Yaralov, G. (2003). Sequential detection of targets in multichannel systems. *IEEE Transactions on Information Theory*, 49(2):425–445.

[74] Tartakovsky, A. G. and Veeravalli, V. V. (2005). General asymptotic bayesian theory of quickest change detection. *Theory of Probability & Its Applications*, 49(3):458–497.

[75] Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

[76] Woodroofe, M. (1982). *Nonlinear renewal theory in sequential analysis*, volume 39. Siam.

[77] Ye, S., Fellouris, G., Culpepper, S., and Douglas, J. (2016). Sequential detection of learning in cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*.

[78] Zhang, S. and Chang, H.-H. (2016). From smart testing to smart learning: how testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 1(1):67–92.