Taylor & Francis
Taylor & Francis Group

Check for updates

Beyond boundaries: Gary Lorden's groundbreaking contributions to sequential analysis

Jay Bartroff[a] and Alexander G. Tartakovsky[b]

[a]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, Texas, USA; [b]AGT StatConsult, Los Angeles, California, USA

ABSTRACT

Gary Lorden provided several fundamental and novel insights into sequential hypothesis testing and change point detection. In this article, we provide an overview of Lorden's contributions in the context of existing results in those areas and some extensions made possible by Lorden's work. We also mention some of Lorden's significant consulting work, including as an expert witness and for NASA, the entertainment industry, and Major League Baseball.

1. INTRODUCTION

The purpose of this article is to provide an overview of Gary Lorden's significant contributions to the field of sequential analysis. But first, to give a sense of Lorden's remarkable life and personality, in Sections 1.1 and 1.2 we give a brief biography of Lorden and highlight some of his extra-academic work.

1.1. Biographical Sketch

Gary Allen Lorden was born in Los Angeles, California, on June 10, 1941. Lorden entered Caltech as a freshman in 1958, and Lorden's undergraduate contemporaries at Caltech included a number of others who would also go on to be notable statisticians, including Larry Brown, Peter Bickel, Brad Efron, and Carl Morris. Lorden received a B.S. from Caltech in 1962 and a Ph.D. from Cornell University in 1966 under the supervision of Jack Kiefer. After a faculty position at Northwestern University, Lorden rejoined Caltech in 1968, where he stayed until his retirement in 2009. Figure 1 shows some photos of Lorden at different stages of his career.

Beyond research and teaching, Lorden was known for his leadership roles at Caltech. He served as dean of students from 1984 to 1988, vice president for student affairs from 1989 to 1998, and acting vice president for student affairs in 2002. He was executive officer (Caltech's version of department chair) for the mathematics department from 2003 to 2006.

CONTACT Jay Bartroff ✉ bartroff@austin.utexas.edu 🖥 Department of Statistics and Data Sciences, University of Texas at Austin, 2317 Speedway, Mail Code D9800, Austin, TX 78712, USA.
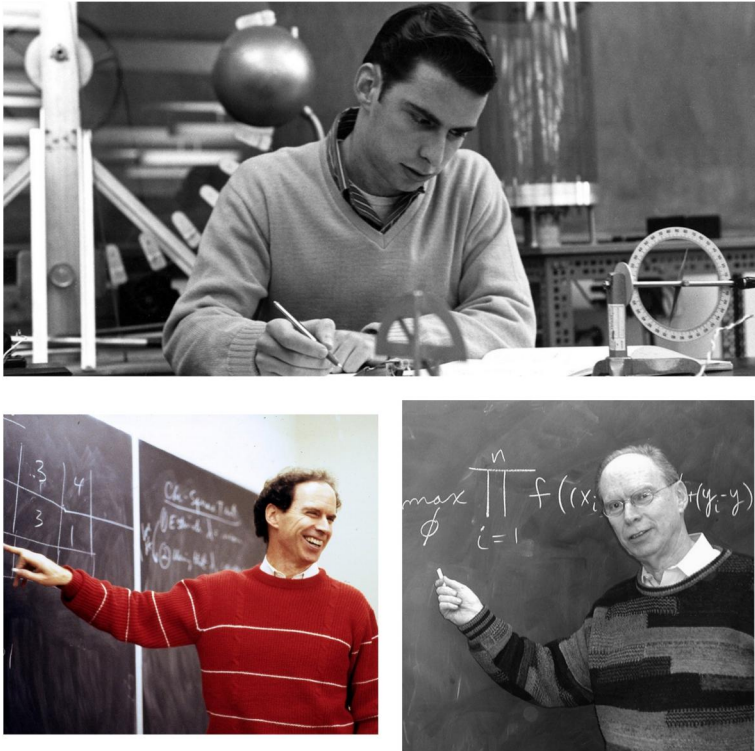
Figure 1. Counterclockwise from top: Gary Lorden as an undergraduate at Caltech in 1959, lecturing in 1987, and lecturing circa 2010. Photos courtesy of Caltech.

Gary and his wife, Louise, were both accomplished pianists and enjoyed playing and singing duets for guests, often students, at their home in Pasadena. They also enjoyed showing students how well they danced together. Lorden also liked to act and regularly participated in plays put on at Caltech.

Lorden passed away on October 25, 2023, at the age of 82. He is survived by his children Lisa and Diana; his wife Louise passed away in 2015.

1.2. Consulting, Hollywood, and Major League Baseball

Lorden was known for his creativity and generosity of ideas, so it is not surprising that throughout his career he was in demand as an academic collaborator and consultant, and much of his early work on change point detection arose from collaborations with researchers at the Jet Propulsion Laboratory in Pasadena, California, which is managed for NASA by Lorden's home institution of Caltech.

But Lorden was also an uncommonly effective, engaging, and entertaining communicator, and this caused many outside academics to seek him out as a consultant, too. Lorden would routinely be interviewed by reporters and appear on the evening news, explaining things like the odds of winning the latest lottery jackpot. Lorden also "moonlighted" as an expert witness in court cases involving statistics and math.

Eventually Hollywood came calling, and in 2005 Lorden was asked to be the math consultant for a (then) new CBS television show called *NUMB3RS*. In a bit of art imitating life, the show was about a Caltech professor who used math to help the FBI solve crimes, and Lorden worked with the show's writers to accurately incorporate mathematical topics into the storylines. An aspect of the relationship that Lorden particularly relished, which he related to us, was that the show's creators initially considered setting the show at MIT (Caltech's rival) but decided to change the venue after learning more about Caltech and Lorden. The show would go on to be a hit, running for 118 episodes over six seasons. If a viewer of the show knew something of Lorden's work, they could see some of his favorite topics (and many topics discussed below in this article) woven into the episodes' plot lines, including change point detection, hypothesis testing, Bayesian methods, gambling math, cryptography, and sports statistics, among others. With Keith Devlin, Lorden wrote a popular general audience book (Devlin and Lorden 2007) on the mathematical topics appearing in the show. For example, Devlin and Lorden (2007, ch. 4) explained the basic concept of change point detection—the determination that a definite change has occurred, as opposed to normal fluctuations—and went on to discuss this method's importance for quick response to potential bioterrorist attacks and for designing efficient algorithms to pinpoint various kinds of criminal activity, such as to detect an increase in crime rates in certain geographical areas and to track changes in financial transactions that could be criminal.

Another high-profile consulting project came in 2018 when Lorden, Bartroff, and others were chosen by the commissioner of Major League Baseball (MLB) to be statisticians on a committee with physicists, engineers, and baseball experts studying MLB's then-recent surge in home runs. The committee studied vast amounts of Statcast game data, as well as laboratory tests on the properties of the baseball that can affect home run production, including Lorden traveling to Costa Rica to inspect baseball manufacturer Rawlings' production plant there. The committee made recommendations (Nathan et al. 2018) to MLB and Rawlings for future monitoring, testing, and storage of baseballs.

1.3. The Remainder of This Article

The remainder of this article is devoted to describing Lorden's fundamental and far-reaching results in sequential hypothesis testing and change point detection, which we aim to present within the context of those areas. Beginning with hypothesis testing in Section 2, after describing the testing setup and optimality of the sequential probability ratio test (SPRT) in Section 2.1.1, Lorden's findings on multiparameter testing and their application to near-optimality of the multihypothesis SPRT are covered in Section 2.1.2. We then cover Lorden's fundamental inequality for excess over the boundary in Section 2.2. Additionally, we explore Lorden's contributions to the Keifer-Weiss problem of testing while minimizing the expected sample size at a parameter value between the hypotheses and other results stemming from his work (Section 2.3), optimal testing of composite hypotheses (Section 2.4), and optimal multistage testing (Section 2.5). In Section 3 we cover Lorden's fundamental minimax change point detection theory and related advancements in the field.

2. SEQUENTIAL HYPOTHESIS TESTING

In this section, we delve into Lorden's contributions to hypothesis testing, encompassing the challenges of excess over the boundaries of random walks, the formulation of nearly optimal multi-decision sequential rules (including multihypothesis tests and multistage tests), and the modified Keifer-Weiss problem. These topics were extensively explored in the seminal Lorden's papers (Lorden 1967, 1970, 1972, 1973, 1976, 1977b, 1980, 1983).

2.1. Multihypothesis Testing

2.1.1. The General Multihypothesis Testing Problem

One of Lorden's major fundamental contributions is the proposal of a multihypothesis sequential test that achieves third-order asymptotic optimality in the independent and identically distributed (i.i.d.) case, akin to Wald's SPRT when error probabilities are small. Further details will be provided in Subsection 2.1.2. Additionally, extensions to a more general non-i.i.d. case, where observations may exhibit dependency and nonidentical distributions, have been explored in Lai (1981), Tartakovsky (1998, 2020, 2024), and Tartakovsky, Nikiforov, and Basseville (2015).

We begin with formulating the following multihypothesis testing problem as addressed by Lorden (1977b). Let $(\Omega, \mathscr{F}, \mathscr{F}_n, \mathsf{P})$, $n \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$, be a filtered probability space, where the sub-σ-algebra $\mathscr{F}_n = \sigma(\mathbf{X}^n)$ of $\mathscr{F}$ is generated by the sequence of random variables $\mathbf{X}^n = \{X_t, \ 1 \leqslant t \leqslant n\}$ observed up to time n, which is defined on the space $(\Omega, \mathscr{F})$ ($\mathscr{F}_0$ is trivial). The focus lies on the N-decision problem of testing the hypotheses $\mathcal{H}_i : \mathsf{P} = \mathsf{P}_i, \ i = 1, \dots, N$, where $\mathsf{P}_1, \dots, \mathsf{P}_N$ are given probability measures assumed to be locally mutually absolutely continuous; that is, their restrictions $\mathsf{P}_i^n = \mathsf{P}|_{\mathscr{F}_n}$ and $\mathsf{P}_j^n = \mathsf{P}_j|_{\mathscr{F}_n}$ to $\mathscr{F}_n$ are equivalent for all $1 \leqslant n < \infty$ and all $i, j = 1, \dots, N, \ i \neq j$. Let Q^n be a restriction to $\mathscr{F}_n$ of a nondegenerate σ-finite measure Q on $(\Omega, \mathscr{F})$.

Assume that the observed random variables $X_1, X_2, \dots$ are i.i.d., so that under P_i the sample $\mathbf{X}^n = (X_1, \dots, X_n)$ has joint density $p_i(\mathbf{X}^n)$ with respect to the dominating measure Q^n for all $n \in \mathbb{N}$, which can be expressed as

$$p_i(\mathbf{X}^n) = \prod_{t=1}^{n} f_i(X_t), \quad i = 1, \dots, N, \tag{2.1}$$

where $f_i(X_t)$ represents the respective density for X_t under the hypothesis $\mathcal{H}_i$. Hence, the interest lies in determining which of N given densities $f_1, \dots, f_N$ is true.

A multihypothesis sequential test is a pair $\delta = (T, d)$, where T is a stopping time with respect to the filtration $\{\mathscr{F}_n\}_{n \in \mathbb{Z}_+}$ and $d = d(\mathbf{X}^T)$ is an $\mathscr{F}_T$-measurable terminal decision function with values in the set $\{1, \dots, N\}$. Specifically, $d = i$ means that the hypothesis $\mathcal{H}_i$ is accepted upon stopping, $\{d = i\} = \{T < \infty, \delta \text{ accepts } \mathcal{H}_i\}$. Let $\alpha_{ij}(\delta) = \mathsf{P}_i(d = j), \ i \neq j$, $i, j = 1, \dots, N$, denote the error probabilities of the test δ; that is, the probabilities of accepting the hypothesis $\mathcal{H}_j$ when $\mathcal{H}_i$ is true.

Introduce the class of tests with probabilities of errors $\alpha_{ij}(\delta)$ that do not exceed the prespecified numbers $0 < \alpha_{ij} < 1$:

$$\mathbb{C}(\boldsymbol{\alpha}) = \left\{\delta : \alpha_{ij}(\delta) \leqslant \alpha_{ij} \text{ for } i, j = 1, \dots, N, i \neq j\right\}, \tag{2.2}$$

where $\boldsymbol{\alpha} = (\alpha_{ij})$ is a matrix of given error probabilities that are positive numbers less than 1 (the diagonal entries α_{ii} are immaterial).

Let E_i denote the expectation under hypothesis $\mathcal{H}_i$ (i.e., under the measure P_i). The objective is to discover a sequential test that would minimize the expected sample sizes $\mathsf{E}_i[T]$ for all hypotheses $\mathcal{H}_i, \ i = 1, ..., N$, at least approximately.

For $n \in \mathbb{N}$, define the likelihood ratio (LR) and the log-likelihood ratio (LLR) processes between hypotheses $\mathcal{H}_i$ and $\mathcal{H}_j$ as

$$\Lambda_{ij}(n) = \frac{\mathrm{d}\mathsf{P}_i^n}{\mathrm{d}\mathsf{P}_j^n}(\mathbf{X}^n) = \frac{p_i(\mathbf{X}^n)}{p_j(\mathbf{X}^n)} = \prod_{t=1}^n \frac{f_i(X_t)}{f_j(X_t)},$$

$$\lambda_{ij}(n) = \log \Lambda_{ij}(n) = \sum_{t=1}^n \log \left[\frac{f_i(X_t)}{f_j(X_t)} \right].$$

In a particular case of two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ ($N = 2$), Wald (1945, 1947) introduced the SPRT. Let $Z_t = \log[f_1(X_t)/f_2(X_t)]$ be the LLR for the observation X_t, so the LLR for the sample $\mathbf{X}^n$ is the sum

$$\lambda_{12}(n) = \lambda_n = \sum_{t=1}^n Z_t, \quad n = 1, 2, ... \tag{2.3}$$

Letting $a_0 < 0$ and $a_1 > 0$ be thresholds, Wald's SPRT $\delta_*(a_0, a_1) = (T_*, d_*)$ is

$$T_*(a_0, a_1) = \inf\{n \geqslant 1 : \lambda_n \notin (a_0, a_1)\}, \quad d_*(a_0, a_1) = \begin{cases} 1 & \text{if } \lambda_{T_*} \geqslant a_1 \\ 2 & \text{if } \lambda_{T_*} \leqslant a_0. \end{cases} \tag{2.4}$$

In the case of two hypotheses, the class of tests (2.2) is defined as

$$\mathbb{C}(\alpha_0, \alpha_1) = \{\delta : \alpha_0(\delta) \leqslant \alpha_0 \ \text{and} \ \alpha_1(\delta) \leqslant \alpha_1\},$$

where α_0 represents the upper bound on the type I error (false positive) probability $\alpha_0(\delta) = \alpha_{12}(\delta)$, and α_1 represents the upper bound on the type II error (false negative) probability $\alpha_1(\delta) = \alpha_{21}(\delta)$.

Wald's SPRT possesses an extraordinary optimality property: it minimizes both the expected sample sizes $\mathsf{E}_1[T]$ and $\mathsf{E}_2[T]$ within the class of sequential (and nonsequential) tests $\mathbb{C}(\alpha_0, \alpha_1)$ as long as the observations are i.i.d. under both hypotheses. In other words, $\mathsf{E}_i[T_*] = \inf_{\delta \in \mathbb{C}(\alpha_0, \alpha_1)} \mathsf{E}_i[T]$ for $i = 0, 1$, as established by Wald and Wolfowitz (1948) through a Bayesian approach.

Lai (1981) proved that the SPRT is also first-order asymptotically optimal as $\max(\alpha_0, \alpha_1) \to 0$ for general non-i.i.d. models with dependent and non–identically distributed observations when the normalized log-likelihood ratio $n^{-1}\lambda_n$ converges 1-quickly to finite numbers I_i under P_i.

The central idea of Lorden's investigation, elaborated in detail in Section 2.1.2, is that, similar to how the SPRT is strictly optimal in the class $\mathbb{C}(\alpha_0, \alpha_1)$ for any error probabilities α_0 and α_1, combinations of SPRTs exhibit third-order asymptotic optimality for multihypothesis testing problems involving any finite number of densities when probabilities of errors are small.

2.1.2. Near-Optimality of the Multihypothesis SPRT

The problem of sequentially testing many hypotheses is substantially more complex than that of testing two hypotheses. Identifying an optimal test in the class (2.2) that minimizes expected sample sizes for all hypotheses $\mathcal{H}_1, ..., \mathcal{H}_N$, is daunting. Hence, a significant portion of the development of sequential multihypothesis testing in the 20th century has focused on the exploration of certain combinations of one-sided SPRTs. See Armitage (1950), Chernoff (1959), Kiefer and Sacks (1963), and Lorden (1967, 1977b).

The results of Lorden's ingenious paper (Lorden 1977b) are of fundamental importance because they establish third-order asymptotic optimality of the accepting multihypothesis test that he proposed. More specifically, Lorden established that just as the SPRT is optimal in the class $\mathbb{C}(\alpha_0, \alpha_1)$ for testing two hypotheses, certain combinations of one-sided SPRTs are nearly optimal in a third-order sense in the class $\mathbb{C}(\boldsymbol{\alpha})$; that is, subject to error probability constraints, expected sample sizes are minimized to within the negligible additive $o(1)$ term:

$$\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{E}_i[T] = \mathsf{E}_i[T_*] + o(1) \quad \text{as } \alpha_{\max} \to 0 \quad \text{for all } i = 1, ..., N, \tag{2.5}$$

where $\alpha_{\max} = \max_{1 \leqslant i, j \leqslant N, i \neq j} \alpha_{ij}$ and T_* is the stopping time of the multihypothesis test δ_*, which is defined below.

We now define a test proposed by Lorden, which we will refer to as the accepting matrix SPRT (MSPRT). Write $\mathcal{N} = \{1, ..., N\}$. For a threshold matrix $\mathbf{A} = (A_{ij})_{i,j \in \mathcal{N}}$, with $A_{ij} > 0$ for $i \neq j$ and the A_{ii} are immaterial (0, say), define the MSPRT $\delta_* = (T_*, d_*)$, built on one-sided SPRTs between hypotheses $\mathcal{H}_i$ and $\mathcal{H}_j$, as follows:

$$\text{Stop at the first } n \geqslant 1 \text{ such that, for some } i, \Lambda_{ij}(n) \geqslant A_{ji} \text{ for all } j \neq i, \tag{2.6}$$

and accept the unique $\mathcal{H}_i$ that satisfies these inequalities. Note that for $N = 2$ the MSPRT coincides with Wald's SPRT.

Let $a_{ji} = \log A_{ji}$. Introducing the Markov accepting times for hypotheses $\mathcal{H}_i$ as

$$T_i = \inf \left\{ n \geqslant 1 : \min_{\substack{1 \leqslant j \leqslant N \\ j \neq i}} \left[\lambda_{ij}(n) - a_{ji} \right] \geqslant 0 \right\}, \quad i = 1, ..., N, \tag{2.7}$$

the test in equation (2.6) can also be written in the following form:

$$T_* = \min_{1 \leqslant j \leqslant N} T_j, \qquad d_* = i \quad \text{if} \quad T_* = T_i. \tag{2.8}$$

Thus, in the MSPRT, each component SPRT is extended until, for some $i \in \mathcal{N}$, all $N - 1$ SPRTs involving $\mathcal{H}_i$ accept $\mathcal{H}_i$.

The MSPRT is not strictly optimal for $N > 2$, but it is a good approximation to the optimal multihypothesis test. Under certain conditions and with some choice of the threshold matrix $\mathbf{A}$, it minimizes the expected sample sizes $\mathsf{E}_i[T]$ for all $i = 1, ..., N$ to within a vanishing $o(1)$ term for small error probabilities; see (2.5).

Consider first the first-order asymptotic criterion: Find a multihypothesis test $\delta_*(\boldsymbol{\alpha}) = (d_*(\boldsymbol{\alpha}), T_*(\boldsymbol{\alpha}))$ such that

$$\lim_{\alpha_{\max} \to 0} \frac{\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{E}_i[T]}{\mathsf{E}_i[T_*(\boldsymbol{\alpha})]} = 1 \quad \text{for all } i = 1, ..., N. \tag{2.9}$$

Using Wald's likelihood ratio identity, it is easily shown that $\alpha_{ij}(\delta_*) \leqslant \exp(-a_{ij})$ for $i,j = 1, ..., N$, $i \neq j$, so selecting $a_{ji} = |\log \alpha_{ji}|$ implies $\delta_* \in \mathbb{C}(\boldsymbol{\alpha})$. These inequalities are similar to Wald's in the binary hypothesis case and are very imprecise. Using Wald's approach, it is rather easy to prove that the MSPRT with boundaries $a_{ji} = |\log \alpha_{ji}|$ is first-order asymptotically optimal, minimizing expected sample sizes as long as the Kullback-Leibler information numbers $I_{ij} = \mathsf{E}_i[\lambda_{ij}(1)]$ are positive and finite; see Tartakovsky, Nikiforov, and Basseville (2015, section 4.3.1).

In his ingenious paper, Lorden (1977b) substantially improved this result, showing that with a sophisticated design that includes accurate estimation of thresholds accounting for overshoots, the MSPRT is nearly optimal in the third-order sense (2.5).

Specifically, assume the second-moment condition

$$\mathsf{E}_i\big[\lambda_{ij}(1)\big]^2 < \infty, \quad i,j = 1, ..., N \tag{2.10}$$

and define the numbers

$$\mathcal{L}_{ij} = \exp\left\{-\sum_{n=1}^{\infty} \frac{1}{n}\big[\mathsf{P}_j(\lambda_{ij}(n) > 0) + \mathsf{P}_i(\lambda_{ij}(n) \leqslant 0)\big]\right\}, \quad i,j = 1, ..., N. \tag{2.11}$$

These numbers are symmetric, $\mathcal{L}_{ij} = \mathcal{L}_{ji}$, and $0 < \mathcal{L}_{ij} \leqslant 1$ ($\mathcal{L}_{ii} \equiv 1$). Furthermore, $\mathcal{L}_{ij} = 1$ only if the measures P_i^n and P_j^n are singular so that the absolute continuity assumption is violated.

For $i,j \in \mathcal{N}$ ($i \neq j$) and $a > 0$, define one-sided SPRTs

$$\tau_{ij}(a) = \inf\big\{n \geqslant 0 : \lambda_{ij}(n) \geqslant a\big\}. \tag{2.12}$$

Using a renewal-theoretic argument, the numbers $\mathcal{L}_{ij}$ are tightly related to the overshoots in the one-sided tests. If the LLR $\lambda_{ij}(1)$ is non-arithmetic under $\mathcal{H}_i$, then

$$\mathcal{L}_{ij} = \zeta_{ij} I_{ij}, \quad \zeta_{ij} = \lim_{a\to\infty} \mathsf{E}_i\big\{\exp\big[-(\lambda_{ij}(\tau_{ij}(a)) - a)\big]\big\} \tag{2.13}$$

(see, e.g., theorem 3.1.3 in Tartakovsky, Nikiforov, and Basseville 2015).

It turns out that the $\mathcal{L}$-numbers play a significant role both in the Bayes and the frequentist frameworks. They facilitate the adjustment of boundaries necessary to achieve optimality.

Consider the Bayes multihypothesis problem with the prior distribution of hypotheses $\boldsymbol{\pi} = (\pi_0(1), ..., \pi_0(N))$, where $\pi_0(i) = \mathsf{P}(\mathcal{H}_i)$, and the loss incurred when stopping at time $T = n$ and making the decision $d = j$ while the hypothesis $\mathcal{H}_i$ is true is $L_n(\mathcal{H}_i, d = j, \mathbf{X}^n) = L_{ij} + cn$, where $c > 0$ is the cost of making one observation or sampling cost and where $0 < L_{ij} < \infty$ for $i \neq j$ and 0 if $i = j$.

The average (integrated) risk of the test $\delta = (T, d)$ is

$$\rho_c^{\pi}(\delta) = \sum_{i=1}^{N} \pi_0(i)\left[\sum_{j=1}^{N} L_{ij}\mathsf{P}_i(d = j) + c\,\mathsf{E}_i[T]\right].$$

It follows from theorem 1 of Lorden (1977b) that, as $c \to 0$, the MSPRT δ_* defined in equation (2.6) with the thresholds $A_{ji}(c) = (\pi_0(j)/\pi_0(i))L_{ji}\mathcal{L}_{ij}/c$ is

asymptotically third-order optimal (i.e., to within $o(c)$) under the second-moment condition (2.10):

$$\rho_c^\pi(\delta^*) = \inf_\delta \ \rho_c^\pi(\delta) + o(c) \quad \text{as } c \to 0,$$

where infimum is taken over all sequential or nonsequential tests.

Using this Bayes asymptotic optimality result, it can be proven that the MSPRT is also nearly optimal to within $o(1)$ with respect to the expected sample sizes $\mathsf{E}_i[T]$ for all hypotheses among all tests with constrained error probabilities. In other words, the MSPRT has an asymptotic property similar to the exact optimality of the SPRT for two hypotheses. This result is more practical than the above Bayes optimality.

The following theorem provides detailed specifications, resembling theorem 4 and its corollary in Lorden (1977b). Recall that $\alpha_{ij}(\delta) = \mathsf{P}_i(d = j)$ represents the probability of erroneously accepting the hypothesis $\mathcal{H}_j$ when $\mathcal{H}_i$ is true. In addition, denote as $\tilde{\alpha}_i(\delta) = \mathsf{P}_i(d \neq i)$ the probability of erroneously rejecting $\mathcal{H}_i$ when it is true and $\beta_j(\delta) = \sum_{i=1}^N w_{ij}\mathsf{P}_i(d = j)$ as the weighted probability of accepting $\mathcal{H}_j$, where $(w_{ij})_{i,j\in\mathcal{N}}$ is a given matrix of positive weights. Recall the definition of the class of tests (2.2) for which the probabilities of errors $\mathsf{P}_i(d = j)$ do not exceed prescribed values α_{ij} and introduce two more classes that upper-bound the weighted probabilities of errors $\beta_j(\delta)$ and probabilities of errors $\tilde{\alpha}_i(\delta)$, respectively:

$$\bar{\mathbb{C}}(\boldsymbol{\beta}) = \big\{\delta : \beta_j(\delta) \leqslant \beta_j \text{ for } j = 1, ..., N\big\}, \tag{2.14}$$

$$\tilde{\mathbb{C}}(\tilde{\boldsymbol{\alpha}}) = \big\{\delta : \tilde{\alpha}_i(\delta) \leqslant \tilde{\alpha}_i \text{ for } i = 1, ..., N\big\}. \tag{2.15}$$

If $A_{ij} = A_{ij}(c)$ is a function of the small parameter c, then the error probabilities $\alpha_{ij}^*(c)$, $\tilde{\alpha}_i^*(c)$, and $\beta_j^*(c)$ of the MSPRT $\delta_*(c)$ are also functions of this parameter, and if $A_{ji}(c) \to \infty$, then $\alpha_{ij}^*(c), \beta_j^*(c) \to 0$ as $c \to 0$. Note that $\tilde{\alpha}_i^*(c) = \sum_{j\neq i} \alpha_{ij}^*(c)$, so it also goes to zero as $c \to 0$. We denote as $\boldsymbol{\beta}^*(c)$ the vector $(\beta_1^*(c), ..., \beta_N^*(c))$, as $\tilde{\boldsymbol{\alpha}}^*(c)$ the vector $(\tilde{\alpha}_1^*(c), ..., \tilde{\alpha}_N^*(c))$, and as $\boldsymbol{\alpha}^*(c)$ the matrix $(\alpha_{ij}^*(c))_{i,j\in\mathcal{N}}$.

Theorem 2.1 (MSPRT near optimality). *Assume that the second moment condition (2.10) holds.*

i. *If the thresholds in the MSPRT are selected as $A_{ji}(c) = w_{ji}\mathcal{L}_{ij}/c, \ i,j = 1, ..., N$, then*

$$\mathsf{E}_i[T^*(c)] = \inf_{\delta\in\bar{\mathbb{C}}(\boldsymbol{\beta}^*(c))} \mathsf{E}_i[T] + o(1) \quad \text{as } c \to 0 \text{ for all } i = 1, ..., N, \tag{2.16}$$

that is, the MSPRT minimizes to within $o(1)$ the expected sample sizes among all tests whose weighted error probabilities are less than or equal to those of $\delta^*(c)$.

ii. *For any matrix $\mathbf{B} = (B_{ij}) \ (B_{ij} > 0, \ i \neq j)$, let $A_{ji} = B_{ji}/c$. The MSPRT $\delta^*(c)$ asymptotically minimizes the expected sample sizes for all hypotheses to within $o(1)$ as $c \to 0$ among all tests whose error probabilities $\alpha_{ij}(\delta)$ are less than or equal to those of $\delta^*(c)$ as well as whose error probabilities $\tilde{\alpha}_i(\delta)$ are less than or equal to those of $\delta^*(c)$; that is,*

$$\mathsf{E}_i[T^*(c)] = \inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha}^*(c))} \mathsf{E}_i[T] + o(1) \quad \text{as } c \to 0 \quad \text{for all } i = 1, ..., N \qquad (2.17)$$

and

$$\mathsf{E}_i[T^*(c)] = \inf_{\delta \in \tilde{\mathbb{C}}(\tilde{\boldsymbol{\alpha}}^*(c))} \mathsf{E}_i[T] + o(1) \quad \text{as } c \to 0 \quad \text{for all } i = 1, ..., N. \qquad (2.18)$$

The intuition behind these results is that because the MSPRT is a combination of one-sided SPRTs $\tau_{ij}(a_{ji})$ defined in equation (2.12) and because the $\zeta_{ij} = \mathcal{L}_{ij}/I_{ij}$ are correction factors to the error probability bound $\mathsf{P}_j(\tau_{ij}(a_{ji}) < \infty) \leqslant e^{-a_{ji}}$, the asymptotic approximation

$$\mathsf{P}_j(\tau_{ij}(a_{ji}) < \infty) = \zeta_{ij} e^{-a_{ji}}(1 + o(1)) \quad \text{as } a_{ji} \to \infty,$$

works well even for moderate values of a_{ji}. So taking $a_{ji} = \log(I_{ij}/\mathcal{L}_{ij}\alpha)$ allows one to attain a nearly optimal solution in the frequentist problem. The proofs of these results are extremely tedious and require many nonstandard and sophisticated mathematical tools developed by Lorden.

Notice that Theorem 2.1 only addresses the asymptotically symmetric case where

$$\lim_{c \to 0} \frac{\log \beta_j^*(c)}{\log \beta_k^*(c)} = 1, \quad \lim_{c \to 0} \frac{\log \tilde{\alpha}_i^*(c)}{\log \tilde{\alpha}_k^*(c)} = 1 \quad \text{and} \quad \lim_{c \to 0} \frac{\log \alpha_{ij}^*(c)}{\log \alpha_{ks}^*(c)} = 1. \qquad (2.19)$$

Introducing for the hypotheses $\mathcal{H}_i$ different observation costs c_i that may go to 0 at different rates—that is, setting $A_{ji} = B_{ji}/c_i$—the results of Theorem 2.1 can be generalized to the more general asymmetric case where the ratios in equation (2.19) are bounded away from zero and infinity. This generalization is important for certain applications.

Lorden's outlined results and methodologies hold significant potential for application across various problems and domains. For instance, consider their relevance in the multistream (or multichannel) problem involving two decisions and multiple data streams, as explored by Fellouris and Tartakovsky (2017) and discussed in Tartakovsky (2020, ch. 1). Sequential hypothesis testing within multiple data streams, such as sensors, populations, or multichannel systems, carries numerous practical implications and applications.

Suppose that observations are sequentially acquired over time in N streams. The observations in the ith data stream correspond to a realization of a stochastic process $X(i) = \{X_n(i)\}_{n \in \mathbb{N}}$, where $i \in \mathcal{N} := \{1, ..., N\}$ and $\mathbb{N} = \{1, 2, ...\}$. Let $\mathcal{H}_0$ be the null hypothesis according to which all N streams are not affected; that is, there are no "signals" in all streams at all. For any given non-empty subset of components, $\mathcal{B} \subset \mathcal{N}$, let $\mathcal{H}_{\mathcal{B}}$ be the hypothesis according to which only the components $X(i)$ with i in $\mathcal{B}$ contain signals. Denote by P_0 and $\mathsf{P}_{\mathcal{B}}$ the distributions of $\mathbf{X}$ under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_{\mathcal{B}}$, respectively. Next, let $\mathcal{P}$ be a class of subsets of $\mathcal{N}$ that incorporates a priori information that may be available regarding the subset of affected streams. Denote by $|\mathcal{B}|$ the size of a subset $\mathcal{B}$—that is, the number of signals under $\mathcal{H}_{\mathcal{B}}$—and by $|\mathcal{P}|$ the size of class $\mathcal{P}$; that is, the number of possible alternatives in $\mathcal{P}$. For example, if we know upper $\bar{K} \leqslant N$ and lower $\underline{K} \geqslant 1$ bounds on the size of the affected subset or when we

know that at most K streams can be affected, then $\mathcal{P} = \mathcal{P}_{\underline{K},\bar{K}} = \{\mathcal{B} \subset \mathcal{N} : \underline{K} \leqslant |\mathcal{B}| \leqslant \bar{K}\}$ and $\mathcal{P} = \mathcal{P}_K = \{\mathcal{B} \subset \mathcal{N} : 1 \leqslant |\mathcal{B}| \leqslant K\}$, respectively.

We aim to test $\mathcal{H}_0$, the simple null hypothesis indicating no signals in any data stream, against the composite alternative $\mathcal{H}_1$, according to which the subset of streams with signals belongs to $\mathcal{P}$. We denote $\mathsf{P}_0^n = \mathsf{P}_0|_{\mathscr{F}_n}$ and $\mathsf{P}_\mathcal{B}^n = \mathsf{P}_\mathcal{B}|_{\mathscr{F}_n}$ as restrictions of probability measures P_0 and $\mathsf{P}_\mathcal{B}$ to the σ-algebra $\mathscr{F}_n$ and let $p_0(\mathbf{X}^n)$ and $p_\mathcal{B}(\mathbf{X}^n)$ denote the corresponding probability densities of these measures with respect to some nondegenerate σ-finite measure, where $\mathbf{X}^n = (\mathbf{X}_1, ..., \mathbf{X}_n)$ denotes the concatenation of the first n observations from all data streams.

In what follows, we confine ourselves to the i.i.d. scenario where observations across streams are independent. Moreover, within specific streams, observations are also independent, possessing densities $g_i(x)$ and $f_i(x)$ if the ith stream is unaffected and contains a signal, respectively. Hence, the hypothesis testing problem can be formulated as

$$\mathcal{H}_0 : \quad p(\mathbf{X}^n) = p_0(\mathbf{X}^n) = \prod_{i=1}^{N} \prod_{t=1}^{n} g_i(X_t(i));$$

$$\mathcal{H}_1 = \bigcup_{\mathcal{B} \in \mathcal{P}} \mathcal{H}_\mathcal{B} : \quad p_\mathcal{B}(\mathbf{X}^n) = \prod_{i \in \mathcal{B}} \prod_{t=1}^{n} f_i(X_t(i)) \times \prod_{i \in \mathcal{N} \setminus \mathcal{B}} \prod_{t=1}^{n} g_i(X_t(i)).$$

Because the hypothesis testing problem is binary, the terminal decision d takes two values 0 and 1, so $d \in \{0, 1\}$ is an $\mathscr{F}_T$-measurable random variable such that $\{d = j\} = \{T < \infty, \mathcal{H}_j \text{ is selected}\}$, $j = 0, 1$.

A sequential test should be designed in such a way that the type I (false alarm) and type II (missed detection) error probabilities are controlled; that is, do not exceed given, user-specified levels. Denote by $\mathbb{C}_\mathcal{P}(\alpha_0, \alpha_1)$ the class of sequential tests with the probability of false alarm below $\alpha_0 \in (0, 1)$ and the probability of missed detection below $\alpha_1 \in (0, 1)$; that is,

$$\mathbb{C}_\mathcal{P}(\alpha_0, \alpha_1) = \left\{ \delta : \mathsf{P}_0(d = 1) \leqslant \alpha_0 \text{ and } \max_{\mathcal{B} \in \mathcal{P}} \mathsf{P}_\mathcal{B}(d = 0) \leqslant \alpha_1 \right\}. \quad (2.20)$$

In general, it is not possible to design the tests that are third-order (to within $o(1)$) or even second-order (to within a constant term $O(1)$) asymptotically optimal as $\alpha_{\max} = \max(\alpha_0, \alpha_1) \to 0$. Only finding a test T_* that minimizes the expected sample sizes $\mathsf{E}_0[T]$ and $\mathsf{E}_\mathcal{B}[T]$ for every $\mathcal{B} \in \mathcal{P}$ to first order is possible; that is,

$$\mathsf{E}_0[T_*] \sim \inf_{\delta \in \mathbb{C}_\mathcal{P}(\alpha_0, \alpha_1)} \mathsf{E}_0[T],$$

$$\mathsf{E}_\mathcal{B}[T_*] \sim \inf_{\delta \in \mathbb{C}_\mathcal{P}(\alpha_0, \alpha_1)} \mathsf{E}_\mathcal{B}[T] \quad \text{for all } \mathcal{B} \in \mathcal{P},$$

where E_0 and $\mathsf{E}_\mathcal{B}$ are expectations under P_0 and $\mathsf{P}_\mathcal{B}$, respectively.

Hereafter we use the notation $x_\alpha \sim y_\alpha$ as $\alpha \to 0$ when $\lim_{\alpha \to 0}(x_\alpha/y_\alpha) = 1$.

Let $\mathcal{P}$ be an arbitrary class of subsets of $\mathcal{N}$. For any $\mathcal{B} \in \mathcal{P}$, let $\Lambda_\mathcal{B}(n)$ be the likelihood ratio of $\mathcal{H}_\mathcal{B}$ against $\mathcal{H}_0$ given the observations from all streams up to time n, and let $\lambda_\mathcal{B}(n)$ be the corresponding LLR,

$$\Lambda_{\mathscr{B}}(n) = \frac{dP_{\mathscr{B}}^n}{dP_0^n} = \prod_{i \in \mathscr{B}} \prod_{t=1}^{n} \frac{f_i(X_t(i))}{g_i(X_t(i))},$$

$$\lambda_{\mathscr{B}}(n) = \log \Lambda_{\mathscr{B}}(n) = \sum_{i \in \mathscr{B}} \sum_{t=1}^{n} \log \left[\frac{f_i(X_t(i))}{g_i(X_t(i))} \right].$$

The natural popular statistic for testing $\mathcal{H}_0$ against $\mathcal{H}_1$ at time n is the maximum (generalized) likelihood ratio (GLR) statistic $\hat{\Lambda}(n) = \max_{\mathscr{B} \in \mathcal{P}} \Lambda_{\mathscr{B}}(n)$. However, applying the conventional GLR statistic leads only to the first-order asymptotically optimal test. To obtain second- and third-order optimality, we need to modify the GLR statistic into the weighed GLR $\hat{\Lambda}(n; \boldsymbol{\pi}) = \max_{\mathscr{B} \in \mathcal{P}} \pi_{\mathscr{B}} \Lambda_{\mathscr{B}}(n)$, where $\boldsymbol{\pi} = \{\pi_{\mathscr{B}}, \mathscr{B} \in \mathcal{P}\}$ is a probability mass function on $\mathcal{N}$ fully supported on $\mathcal{P}$; that is, $\pi_{\mathscr{B}} > 0$ for all $\mathscr{B} \in \mathcal{P}$ and $\sum_{\mathscr{B} \in \mathcal{P}} \pi_{\mathscr{B}} = 1$. The corresponding weighted generalized log-likelihood ratio statistic is $\hat{\lambda}(n; \boldsymbol{\pi}) = \max_{\mathscr{B} \in \mathcal{P}}(\lambda_{\mathscr{B}}(n) + \log \pi_{\mathscr{B}})$.

The generalized sequential likelihood ratio test (GSLRT) $\hat{\delta} = (\hat{T}, \hat{d})$ is defined as

$$\hat{T} = \inf\{n \geqslant 1 : \hat{\lambda}(n; \boldsymbol{\pi}_1) \geqslant a_1 \text{ or } \hat{\lambda}(n; \boldsymbol{\pi}_0) \leqslant -a_0\}, \quad \hat{d} = \begin{cases} 1 & \text{if } \hat{\lambda}(\hat{T}; \boldsymbol{\pi}_1) \geqslant a_1 \\ 0 & \text{if } \hat{\lambda}(\hat{T}; \boldsymbol{\pi}_0) \leqslant -a_0 \end{cases},$$

where $\boldsymbol{\pi}_j = \{\pi_{j, \mathscr{B}}, \mathscr{B} \in \mathcal{P}\}$, $j = 0, 1$ are not necessarily identical weights and $a_0, a_1 > 0$ are thresholds that should be selected appropriately to guarantee the desired error probabilities; that is, so that $\hat{T}$ belongs to class $\mathbb{C}_{\mathcal{P}}(\alpha_0, \alpha_1)$ for given α_0 and α_1 with almost exact equalities. The LLR in the ith stream is $\lambda_i(n) = \sum_{t=1}^{n} \log[f_i(X_n(i))/g_i(X_n(i))]$, so that $\lambda_{\mathscr{B}}(n) = \sum_{i \in \mathscr{B}} \lambda_i(n)$.

The $\mathcal{L}$-number is

$$\mathcal{L}_{\mathscr{B}} = \exp \left\{ -\sum_{n=1}^{\infty} \frac{1}{n} [P_0(\lambda_{\mathscr{B}}(n) > 0) + P_{\mathscr{B}}(\lambda_{\mathscr{B}}(n) \leqslant 0)] \right\}, \tag{2.21}$$

which takes into account the overshoot; compare with Lorden's $\mathcal{L}$-numbers (2.11).

Denote by $\hat{\delta}_*(\boldsymbol{\pi}) = (\hat{T}_*(\boldsymbol{\pi}), \hat{d}_*(\boldsymbol{\pi}))$ the GSLRT with weights

$$\pi_{1, \mathscr{B}} = \frac{\pi_{\mathscr{B}}}{\mathcal{L}_{\mathscr{B}} \sum_{\mathscr{B} \in \mathcal{P}}(\pi_{\mathscr{B}}/\mathcal{L}_{\mathscr{B}})} \quad \text{and} \quad \pi_{0, \mathscr{B}} = \frac{\pi_{\mathscr{B}} \mathcal{L}_{\mathscr{B}}}{\sum_{\mathscr{B} \in \mathcal{P}}(\pi_{\mathscr{B}} \mathcal{L}_{\mathscr{B}})}, \quad \mathscr{B} \in \mathcal{P}. \tag{2.22}$$

The next theorem states that $\hat{\delta}_*(\boldsymbol{\pi})$ is third-order asymptotically optimal, minimizing the weighted expected sample size $E^{\pi}[T]$ to within an $o(1)$ term, where E^{π} is expectation with respect to the probability measure $P^{\pi} = \sum_{\mathscr{B} \in \mathcal{P}} \pi_{\mathscr{B}} P_{\mathscr{B}}$; that is, the weighted expectation $E^{\pi}[\cdot] = \sum_{\mathscr{B} \in \mathcal{P}} \pi_{\mathscr{B}} E_{\mathscr{B}}[\cdot]$.

Theorem 2.2. *Assume the second-moment conditions for LLRs $E_i|\lambda_i(1)|^2 < \infty$ and $E_0|\lambda_i(1)|^2 < \infty$, $i = 1, ..., N$. Let α_0 and α_1 approach 0 so that $|\log \alpha_0|/|\log \alpha_1| \to 1$. If thresholds a_0 and a_1 are selected so that $\hat{\delta}_*(\boldsymbol{\pi})$ belongs to $\mathbb{C}_{\mathcal{P}}(\alpha_0, \alpha_1)$, $P_0(\hat{d}_*(\boldsymbol{\pi}) = 1) \sim \alpha_0$, and $P_1(\hat{d}_*(\boldsymbol{\pi}) = 0) \sim \alpha_1$, then the GSLRT is asymptotically optimal to third order in the class $\mathbb{C}_{\mathcal{P}}(\alpha_0, \alpha_1)$:*

$$\inf_{\delta \in \mathbb{C}_{\mathcal{P}}(\alpha_0, \alpha_1)} \mathsf{E}^{\pi}[T] = \mathsf{E}^{\pi}\big[\hat{T}_*(\boldsymbol{\pi})\big] + o(1) \quad \text{as} \quad \alpha_{\max} \to 0.$$

The central idea of the proof of this result is to consider a purely Bayesian sequential testing problem with the $1 + |\mathcal{P}|$ states "$\mathcal{H}_0$: density g_i for all $i = 1, ..., N$" and "$\mathcal{H}_1^{\mathscr{B}}$: density $f_{\mathscr{B}}$ for $\mathscr{B} \in \mathcal{P}$" and two terminal decisions $d = 0$ (accept $\mathcal{H}_0$) and $d = 1$ (accept $\mathcal{H}_1 = \cup_{\mathscr{B} \in \mathcal{P}} H_1^{\mathscr{B}}$). Then we can exploit Lorden's methods and results to get the proof. Without Lorden's (1977b) paper, this would not be possible. Moreover, the whole idea of using $\mathcal{L}$-numbers for corrections is based on Lorden's fundamental contribution to the field.

2.2. Lorden's (1970) Inequality for the Excess over the Boundary

Partially motivated by seeking improved estimates of the error probabilities and other operating characteristics of Wald's SPRT discussed above, Lorden (1970) considered an upper bound for estimating a random walk's "worst-case" expected overshoot,

$$\sup_{a \geqslant 0} \mathsf{E}[R_a], \tag{2.23}$$

where $a \geqslant 0$ is the boundary,

$$R_a = S_{T(a)} - a \quad \text{is the overshoot,} \tag{2.24}$$

$S_n = \sum_{t=1}^{n} Z_t$ is the random walk, $T(a) = \inf\{n \geqslant 1 : S_n > a\}$ is the stopping time, and, relaxing slightly our notation from Section 2.1.1, here the Z_n are i.i.d. random variables with positive mean m; let Z denote a variate with the same distribution as the Z_n. Wald's (1946) equation tells us that, whenever the following quantities are finite, $m\mathsf{E}[T(a)] = \mathsf{E}[S_{T(a)}] = a + \mathsf{E}[R_a]$, so an upper bound on $\mathsf{E}[R_a]$ provides an upper bound on the expected stopping time $\mathsf{E}[T(a)]$ for the random walk S_n to cross the boundary a. This is closely related to estimates of the expected stopping time $\mathsf{E}[T_*]$ of the SPRT in equation (2.4), as we shall see below.

Wald (1947) provided the upper bound for equation (2.23) of $\sup_{a \geqslant 0} \mathsf{E}[Z - a | Z > a]$, which is exact for the exponential distribution and provides reasonable bounds in some other cases but has serious deficiencies in general: it can be difficult to calculate, is overly conservative in cases like when the distribution of Z has large "gaps," and may be infinite even when $\mathsf{E}[(Z^+)^2] < \infty$, a sufficient condition for finiteness of equation (2.23). Here and throughout this section, $z^+ = \max\{z, 0\}$ is the positive part of z.

For nonnegative Z, results from renewal theory (see Feller 1966) provide estimates of $\mathsf{E}[R_a]$ close to $\mathsf{E}[Z^2]/m = \mathsf{E}[(Z^+)^2]/m$ for both $a = 0$ and as $a \to \infty$. Lorden showed that this is indeed an upper bound for equation (2.23) more generally: for arbitrary i.i.d. Z_n allowed to be discrete or continuous, and taking both positive and negative values, a necessary generalization of the renewal theory results for application to sequential testing and change point detection and analysis in which the Z_n are log-likelihood summands or other sequential test statistic terms.

Theorem 2.3 (Lorden 1970, theorem 1). *If* $Z, Z_1, Z_2, ...,$ *are i.i.d. random variables with mean* $\mathsf{E}[Z] > 0$ *and* $\mathsf{E}[(Z^+)^2] < \infty$*, then* R_a *as defined in equation* (2.24) *satisfies*

$$\sup_{a \geqslant 0} \mathsf{E}[R_a] \leqslant \frac{\mathsf{E}\big[(Z^+)^2\big]}{\mathsf{E}[Z]}. \tag{2.25}$$

Lorden's proof of this theorem involves a number of characteristically clever techniques, of which we highlight a few here. First, he considered the stochastic process $a \mapsto R_a$, noting that (w.p.1) it is piecewise-linear, each "piece" having slope -1. Next, because $a \mapsto R_a$ and even $a \mapsto \mathsf{E}[R_a]$ can behave erratically and be resistant to estimation and bounding, Lorden used the smoothing technique instead of estimating $\int_0^b \mathsf{E}[R_a]\mathrm{d}a$ for $b \geqslant 0$, which is more regularly behaved, as Lorden showed. Finally, the smoothed expected overshoot $\int_0^b \mathsf{E}[R_a]\mathrm{d}a$ is bounded from above using properties of the process R_a and then bounded from below using the following subadditivity property of the integrand $a \mapsto \mathsf{E}[R_a]$ established from the subadditivity of $a \mapsto \mathsf{E}[T(a)]$ and Wald's equation: For any $0 \leqslant a \leqslant b$,

$$\begin{aligned}
\mathsf{E}[R_a] + \mathsf{E}[R_{b-a}] &= \mathsf{E}\big[S_{T(a)}\big] - a + \mathsf{E}\big[S_{T(b-a)}\big] - (b-a) \\
&= m\mathsf{E}[T(a)] + m\mathsf{E}[T(b-a)] - b \quad \text{(Wald's equation)} \\
&\geqslant m\mathsf{E}[T(b)] - b \quad \text{(sub-additivity of } \mathsf{E}[T(b)]) \\
&= \mathsf{E}\big[S_{T(b)}\big] - b \quad \text{(Wald's equation)} \\
&= \mathsf{E}[R_b].
\end{aligned}$$

Returning to the stopping time T_* of the SPRT in equation (2.4), now let the Z_n be the log-likelihood ratio terms as in equation (2.3) and a_0 and a_1 be the boundaries in equation (2.4), and expectation and probability are under the alternative hypothesis density f_1. The random walk S_n now coincides with the log-likelihood ratio statistic λ_n in equation (2.3), although we continue to use the S notation here for clarity. To relate T_* to $T(a_1)$ Lorden observed that

$$S_{T_*} \leqslant \min\{S_{T_*}, a_1\} + (S_{T(a_1)} - a_1), \tag{2.26}$$

and then applying equation (2.25) to the latter term gives the upper bound

$$\mathsf{E}[T_*] \leqslant \frac{(1-\alpha_1)a_1 - \alpha_1 a_0}{m} + \frac{\mathsf{E}\big[(Z^+)^2\big]}{m^2} \tag{2.27}$$

on the expected stopping time of the SPRT under the alternative hypothesis, with a bound under the null hypothesis obtained analogously. Wald (1947) provided a well-known upper bound on the type II error probability α_1, but in order to apply equation (2.27) what is needed is clearly a lower bound on α_1 and a lower bound on α_0 for the corresponding bound under the null. Both of these can be obtained by another application of Lorden's theorem, as follows. Wald's argument gives that

$$\frac{\alpha_0}{1 - \alpha_1} = \mathsf{E}\big[\exp(-S_{T_*})|S_{T_*} > a_1\big].$$

Using the conditional Jensen's inequality with a bound like equation (2.26) after multiplying by the indicator of the event $\{S_{T_*} > a_1\}$, Lorden obtained

$$\frac{\alpha_0}{1 - \alpha_1} \geqslant \exp\left[-\mathsf{E}\big[(S_{T_*}|S_{T_*} > a_1)\big]\right] \geqslant \exp\left[-\left(a_1 + \frac{\mathsf{E}\big[(Z^+)^2\big]}{(1-\alpha_1)m}\right)\right].$$

Using the standard upper bound $\alpha_1 \leqslant e^{-a_1}$, this gives

$$\alpha_0 \geqslant (1 - e^{-a_1}) \exp\left[-\left(a_1 + \frac{\mathsf{E}\left[(Z^+)^2\right]}{(1 - e^{-a_1})m}\right)\right],$$

with an analogous lower bound for α_1.

Lorden (1970, section 2) also obtained generalizations of equation (2.25) to cases in which the variates Z_n are not necessarily i.i.d. They key property is the subadditivity of $T(a)$, for which Lorden assumed the sufficient condition

$$\mathsf{E}\left[(Z_n^+)^2 | T(a) \geqslant n\right] \leqslant r \cdot \mathsf{E}[Z_n | T(a) \geqslant n]$$

for some factor r. Under this condition, Lorden obtained analogous bounds on $\mathsf{E}[R_a]$ and bounds on the moments $\sup_{a \geqslant 0} \mathsf{E}[(R_a)^p]$ for non-i.i.d. observations Z_n (Lorden 1970, theorems 2 and 3), as well as bounds on the tail probability $\mathsf{P}(R_a > x)$ for i.i.d. observations (Lorden 1970, theorem 4).

Other than his seminal 1971 paper on change point detection, Lorden (1970) is his most highly cited paper. In addition to its uses in sequential testing, change point detection, and renewal theory, it has found applications in reliability theory (Rausand and Hoyland 2003), clinical trial design (Whitehead 1997), finance (Novak 2011), and queuing theory (Kalashnikov 2013), among other applications. Perhaps reflecting its fundamental nature and wealth of applications, Lorden's inequality—as equation (2.25) has become known—even has its own Wikipedia entry (https://en.wikipedia.org/wiki/Lorden/%27s_inequality)

2.3. Lorden's 2-SPRT and the Kiefer-Weiss Minimax Optimality

Suppose that based on a sequence of independent observations $\{X_n\}_{n \geqslant 1}$ with common parametric density f_θ one wishes to test the hypothesis $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta = \theta_1$ ($\theta_0 < \theta_1$) with error probabilities at most α_0 and α_1. Even though the SPRT has the remarkable optimality property of minimizing the expected sample size for both statistical hypotheses $\mathsf{E}_{\theta_i}[T]$, $i = 0, 1$, its performance may be poor when the true parameter value $\theta = \vartheta \in (\theta_0, \theta_1)$ differs from putative ones θ_0 or θ_1. Its expected sample size $\mathsf{E}_\vartheta[T]$ can be even much larger than that of the fixed sample size of the Neyman-Pearson test; see, for example, section 5.2 in Tartakovsky, Nikiforov, and Basseville (2015). Much work has been directed toward finding sequential tests that reduce the expected sample size of the SPRT for parameter values between the hypotheses.

Let $\mathbb{C}(\alpha_0, \alpha_1) = \{\delta : \alpha_i(\delta) \leqslant \alpha_i, i = 0, 1\}$ denote the class of tests with error probabilities at most α_0 and α_1 and let

$$\mathsf{ESS}(\alpha_0, \alpha_1) = \inf_{\delta \in \mathbb{C}(\alpha_0, \alpha_1)} \sup_\theta \mathsf{E}_\theta[T]$$

denote the expected sample size of an optimal test in the class $\mathbb{C}(\alpha_0, \alpha_1)$ in the worst-case scenario. The problem of finding a test $\delta_0 = (T_0, d_0)$ such that $\sup_\theta \mathsf{E}_\theta[T_0] = \mathsf{ESS}(\alpha_0, \alpha_1)$ subject to the error probability constraints α_0 and α_1 is known as the Kiefer-Weiss problem. No strictly optimal test has been found so far. Kiefer and Weiss (1957) presented structured results about tests that minimize the expected sample size

$\mathsf{E}_\theta[T]$ at a selected point $\theta = \vartheta \in (\theta_0, \theta_1)$, which is referred to as the modified Kiefer-Weiss problem. Weiss (1962) proved that the Kiefer-Weiss problem reduces to the modified problem in symmetric cases for normal and binomial distributions. Lorden (1976) made a valuable contribution to the modified Kiefer-Weiss problem for two not necessarily parametric hypotheses $\mathcal{H}_i : \mathsf{P} = \mathsf{P}_i$, $i = 0, 1$, when the observations $X_1, X_2, ...$ are i.i.d. and their true distribution P_2 may be different from P_0 and P_1. Lorden (1976) introduced a simple combination of one-sided SPRTs, called the 2-SPRT, and proved that it is third-order asymptotically optimal. Later, Lorden (1980) proved theorems that characterize the basic structure of optimal sequential tests for the modified Kiefer-Weiss problem. His work has generated several works related to both the modified Kiefer-Weiss problem and the original Kiefer-Weiss problem of minimizing the maximal expected sample size; see, for example, Huffman (1983), Dragalin and Novikov (1987), and Tartakovsky, Nikiforov, and Basseville (2015, section 5.3).

Consider the following modified Kiefer-Weiss problem. Let $(\Omega, \mathscr{F}, \mathscr{F}_n, \mathsf{P})$, $n \in \mathbb{Z}_+$, be a filtered probability space where the sub-σ-algebra $\mathscr{F}_n = \sigma(\mathbf{X}^n)$ of $\mathscr{F}$ is generated by the observations $\mathbf{X}^n = \{X_t, \ 1 \leqslant t \leqslant n\}$. The goal is to test the hypotheses $\mathcal{H}_i : \mathsf{P} = \mathsf{P}_i$, $i = 0, 1$, where $\mathsf{P}_0, \mathsf{P}_1$ are given probability measures that are locally mutually absolutely continuous. The true probability measure is either one of P_i or an "intermediate" measure P_2 that is also locally absolute continuous with respect to P_i. Let Q^n be a dominating measure. The observations are i.i.d. under $\mathsf{P}_0, \mathsf{P}_1, \mathsf{P}_2$, so the sample $\mathbf{X}^n = (X_1, ..., X_n)$ has joint densities $p_i(\mathbf{X}^n) = \prod_{t=1}^n f_i(X_t)$ for $i = 0, 1, 2$, with respect to Q^n, where $f_i(X_t)$, $t \geqslant 1$, are densities for the tth observation.

For $n \in \mathbb{N}$ and $i = 0, 1$, define the LR and LLR processes

$$\Lambda_i(n) = \frac{d\mathsf{P}_2^n}{d\mathsf{P}_i^n}(\mathbf{X}^n) = \prod_{t=1}^n \frac{f_2(X_t)}{f_i(X_t)}, \quad \lambda_i(n) = \log \Lambda_i(n) = \sum_{t=1}^n \log \left[\frac{f_2(X_t)}{f_i(X_t)} \right],$$

with $\Lambda_i(0) = 1$ and $\lambda_i(0) = 0$.

Define two parallel one-sided SPRTs:

$$T_0 = \inf\{n \geqslant 1 : \lambda_1(n) \geqslant a_1\}, \quad T_1 = \inf\{n \geqslant 1 : \lambda_0(n) \geqslant a_0\}. \tag{2.28}$$

The stopping time of Lorden's 2-SPRT (Lorden 1976) is $T^\star = \min(T_0, T_1)$ and the terminal decision is $d^\star = \operatorname{argmin}_{i=0,1} T_i$. If $a_i = \log(1/\alpha_i)$, $i = 0, 1$, then $\alpha_i(\delta^\star) = \mathsf{P}_i(d^\star \neq i) \leqslant \alpha_i$; that is, this test belongs to class $\mathbb{C}(\alpha_0, \alpha_1) = \{\delta : \alpha_0(\delta) \leqslant \alpha_0, \alpha_1(\delta) \leqslant \alpha_1\}$. These upper bounds may be rather conservative. For example, in the symmetric case $\mathsf{P}_2(d^\star = 1) = \mathsf{P}_2(d^\star = 0) = 1/2$, we have $\alpha_i(\delta^\star) \leqslant \alpha_i/2$.

Let E_2 denote expectation under P_2, and let $I_i = \mathsf{E}_2[\lambda_i(1)]$, $i = 0, 1$, denote Kullback-Leibler information numbers. The following theorem, proved by Lorden (1976), establishes third-order asymptotic optimality of Lorden's 2-SPRT for small probabilities of errors α_i. Its proof is based on Bayesian arguments. This theorem emerges from theorem 1 in Lorden (1977b), which was proved a year later.

Theorem 2.4. *Let the observations $\{X_n\}_{n \geqslant 1}$ be i.i.d. under P_i, $i = 0, 1, 2$. Assume that the Kullback-Leibler information numbers I_0 and I_1 are positive and, in addition, the second-moment conditions $\mathsf{E}_2|\lambda_i(1)|^2 < \infty$, $i = 0, 1$, hold. Let $\alpha_0^\star(a_0, a_1)$ and $\alpha_1^\star(a_0, a_1)$ denote the error probabilities of the 2-SPRT $\delta^\star(a_0, a_1) = (T^\star(a_0, a_1), d^\star(a_0, a_1))$. Let $\mathsf{ESS}(a_0, a_1)$ denote infimum of the expected sample size $\mathsf{E}_2[T]$ over all tests with $\mathsf{P}_0(d =$*
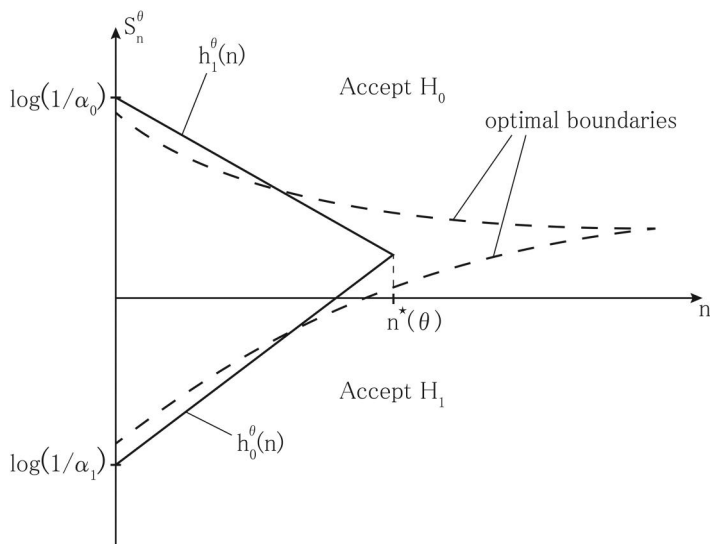
Figure 2. The boundaries $h_1^\theta(n)$ and $h_0^\theta(n)$ of the 2-SPRT (solid) and optimal boundaries (dashed) as functions of n. $S_n^\theta = S_n - nE_\theta[X_1]$.

$1) \leqslant \alpha_0^\star(a_0, a_1)$ *and* $\mathsf{P}_1(d = 0) \leqslant \alpha_1^\star(a_0, a_1)$. *Then*

$$\mathsf{ESS}(a_0, a_1) = \mathsf{E}_2[T^\star(a_0, a_1)] + o(1) \quad \text{as } \min(a_0, a_1) \to \infty, \tag{2.29}$$

where $o(1) \to 0$ *as* $\min(a_0, a_1) \to \infty$.

This theorem implies that if the thresholds a_0 and a_1 in the 2-SPRT are selected so that the error probabilities $\alpha_0^\star(a_0, a_1) = \alpha_0$ and $\alpha_1^\star(a_0, a_1) = \alpha_1$ are exactly equal to the given values α_0 and α_1, then it is third-order asymptotically optimal as $\alpha_{\max} \to 0$ in the class $\mathbb{C}(\alpha_0, \alpha_1)$. The requirement of exact error probabilities can also be relaxed to the asymptotic equalities $\alpha_i^\star(a_0, a_1) = \alpha_i(1 + o(1))$, $i = 0, 1$.

The significance of this result cannot be overstated, because Lorden's simple test is nearly optimal. Simultaneously, the optimal test can be computed using Bellman's backward induction algorithm because the optimal sequential test is truncated, meaning that it has a bounded maximal sample size, as demonstrated by Kiefer and Weiss (1957). For one-parameter exponential families $\{\mathsf{P}_\theta, \theta \in \theta\}$, the optimal bounds exhibit curvature in the (S_n, n) plane, where $S_n = \sum_{t=1}^n X_i$, and determining them typically entails substantial computation. In contrast, Lorden's 2-SPRT approximates optimal curved boundaries with simple linear ones, resulting in a continuation region shaped like a triangle, as illustrated in Figure 2.

Lorden (1976) performed an extensive performance analysis for testing the mean θ of the Gaussian distribution $X_n \sim \mathcal{N}(\theta, 1)$ with the hypotheses $\mathcal{H}_i : \theta = \theta_i$, $(i = 0, 1)$ in the symmetric case where $\alpha_0 = \alpha_1$ and $\mathsf{P}_2 \sim \mathcal{N}(\theta_\star, 1)$, $\theta_\star = (\theta_0 + \theta_1)/2$. The conclusion is that the 2-SPRT performs very closely to the optimal test with its curved boundaries obtained using backward induction. The efficiency depends on the error probabilities, but it was over 99% in all of his performed experiments. Similar results were obtained by Huffman (1983) for the exponential example $f_\theta(x) =$

$\theta e^{-\theta x}$, $x \geqslant 0$, $\theta > 0$. Here the 2-SPRT has efficiency over 98% and almost always over 99% for a broad range of error probabilities and parameter values.

The results of Theorem 2.4 can be extended to the multiple hypothesis case with $N+1$ hypotheses, $N > 1$. Specifically, the modified matrix SPRT is also third-order asymptotically optimal as $\alpha_{\max} = \max_{0 \leqslant i \leqslant N} \alpha_i \to 0$ in the class of tests

$$\mathbb{C}(\boldsymbol{\alpha}) = \left\{ \delta : \alpha_i(\delta) \leqslant \alpha_i \text{ for } i = 0, 1, ..., N \right\},$$

where $\alpha_i(\delta) = \mathsf{P}_i(d \neq i)$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_N)$ is a vector of given error probabilities; cf. Tartakovsky, Nikiforov, and Basseville (2015, theorem 5.3.3, page 240).

Lorden's results have sparked further research into the minimax Kiefer-Weiss problem, which aims to establish near-optimal solutions for the least favorable intermediate distribution P_2 within the single-parameter exponential family. Consider the parametric case $\mathsf{P}_2 = \mathsf{P}_\theta$, $\mathsf{P}_i = \mathsf{P}_{\theta_i}$ where the hypotheses are $\mathcal{H}_0 : \mathsf{P}_0 = \mathsf{P}_{\theta_0}$ and $\mathcal{H}_1 : \mathsf{P}_1 = \mathsf{P}_{\theta_1}$, $\theta_0 < \theta_1$. Let θ be an arbitrary point belonging to the interval (θ_0, θ_1) and let $\delta^\star(\theta) = (d^\star(\theta), T^\star(\theta))$ denote the 2-SPRT tuned to θ. In other words, $T^\star(\theta) = \min(T_0^\theta, T_1^\theta)$, where the T_i^θ's are defined by equation (2.28) with the LLRs $\lambda_i^\theta(n) = \log[d\mathsf{P}_\theta^n / d\mathsf{P}_{\theta_i}^n](\mathbf{X}^n)$, $i = 0, 1$, tuned to θ.

Theorem 2.4 implies that the 2-SPRT $\delta^\star(\theta)$ is third-order asymptotically optimal for minimizing $\mathsf{E}_\theta[T]$ at the intermediate point $\theta \in (\theta_0, \theta_1)$ when the second moments $\mathsf{E}_\theta|\lambda_0^\theta(1)|^2$ and $\mathsf{E}_\theta|\lambda_0^\theta(1)|^2$ are finite and the thresholds a_i are selected so that the error probabilities are either exactly equal, or at least close, to the given numbers α_i, a challenging task. However, setting $a_i = |\log \alpha_i|$ embeds the 2-SPRT into the class $\mathbb{C}(\alpha_0, \alpha_1)$, and Theorem 2.4 suggests that if one can find a nearly least favorable point θ^*—that is, θ^* can be selected so that $\sup_\theta \mathsf{E}_\theta[T^\star(\theta^*)] \approx \mathsf{E}_{\theta^*}[T^\star(\theta^*)]$—then $\delta^\star(\theta^*)$ is an approximate solution to the Kiefer-Weiss problem of minimizing $\sup_\theta \mathsf{E}_\theta[T]$.

For the single-parameter exponential family with density $\{f_\theta(x), \theta \in \Theta\}$, where

$$\frac{f_\theta(x)}{f_{\tilde\theta}(x)} = \exp\left\{ (\theta - \tilde\theta)x - (b(\theta) - b(\tilde\theta)) \right\}, \tag{2.30}$$

with $b(\theta)$ being a convex and infinitely differentiable function on $\tilde\Theta \subset \Theta$, it is feasible to identify the nearly least favorable point $\theta^*(\alpha_0, \alpha_1, \theta_0, \theta_1)$ such that the 2-SPRT with thresholds $a_i = \log(1/\alpha_i)$ achieves second-order asymptotic minimaxity, meaning that the residual term in the discrepancy between the expectation of the sample size of the optimal test and the 2-SPRT is of order $O(1)$ for small α_i. Initially addressed by Huffman (1983), who proposed $\theta^*(\alpha_0, \alpha_1, \theta_0, \theta_1)$ leading to a residual term of order $o(|\log \alpha|^{1/2})$, this problem was further advanced by Dragalin and Novikov (1987), who demonstrated the second-order optimality of Huffman's version of the 2-SPRT.

As previously noted, the formulas $a_i = |\log \alpha_i|$, ensuring the inequalities $\alpha_i(\delta^\star(\theta)) \leqslant \alpha_i$, tend to be overly conservative. A refinement can be achieved by observing that

$$\alpha_1(\delta^\star(\theta)) = \mathsf{P}_\theta(T^\star = T_0^\theta)e^{-a_1}\mathsf{E}_\theta\left\{ e^{-(a_1 - \lambda_1^\theta(T_0^\theta))}|T^\star = T_0^\theta \right\},$$
$$\alpha_0(\delta^\star(\theta)) = \mathsf{P}_\theta(T^\star = T_1^\theta)e^{-a_0}\mathsf{E}_\theta\left\{ e^{-(a_0 - \lambda_0^\theta(T_1^\theta))}|T^\star = T_1^\theta \right\}$$

where, asymptotically as $a_i \to \infty$,

$$\mathsf{E}_\theta \left\{ e^{-(a_0 - \lambda_0^\theta(T_1^\theta))} \big| T^\star = T_1^\theta \right\} \to \zeta_0^\theta, \quad \mathsf{E}_\theta \left\{ e^{-(a_1 - \lambda_1^\theta(T_0^\theta))} \big| T^\star = T_0^\theta \right\} \to \zeta_1^\theta.$$

For the non-arithmetic case, ζ_i^θ can be computed using the renewal-theoretic argument similar to (2.11) and (2.13):

$$\zeta_i^\theta = \frac{1}{I(\theta, \theta_i)} \exp \left\{ -\sum_{n=1}^\infty \frac{1}{n} \left[\mathsf{P}_\theta(\lambda_i^\theta(n) > 0) + \mathsf{P}_i(\lambda_i^\theta(n) \leqslant 0) \right] \right\},$$

where $I(\theta, \theta_i) = (\theta - \theta_i)\dot{b}(\theta) - (b(\theta) - b(\theta_i))$ are the Kullback-Leibler numbers.

The efficiency of Lorden's 2-SPRT in stopping to both reject and accept the null was appealing in clinical trial designs, where these actions are known as efficacy and futility stopping. Lai and Shih (2004) extended the 2-SPRT from the fully sequential setting to the group sequential setting while maintaining its efficiency and balancing the trade-off between efficacy and futility stopping.

2.4. Near-Uniform Optimality of the GLR SPRT for Composite Hypotheses

For practical purposes, it is considerably more significant to devise tests that minimize the expected sample size $\mathsf{E}_\theta[T]$ for all possible parameter values (i.e., uniformly optimal) rather than to address the minimax Kiefer-Weiss problem of minimizing $\mathsf{E}_\theta[T]$ at a least favorable point. In this section, our primary objective is to explore the design of sequential tests that are at least approximately uniformly optimal for small error probabilities or asymptotically Bayesian for a small cost of observations for testing composite hypotheses.

Consider a sequence of i.i.d. observations $X_1, X_2, \ldots,$ originating from a common distribution P_θ with density $f_\theta(x)$ with respect to some nondegenerate sigma-finite measure, where the ℓ-dimensional parameter $\theta = (\theta_1, \ldots, \theta_\ell)$ belongs to a subset Θ of the Euclidean space $\mathbb{R}^\ell$. The parameter space Θ is partitioned into three disjoint sets $\Theta_0, \Theta_1,$ and I_{in}; that is, $\Theta = \Theta_0 \cup \Theta_2 \cup \mathsf{I}_{\text{in}}$. The objective is to test the two composite hypotheses $\mathcal{H}_0 : \theta \in \Theta_0$ against $\mathcal{H}_1 : \theta \in \Theta_1$. The subset I_{in} of Θ denotes an indifference zone where the loss $L(\theta, d)$ associated with correct or incorrect decisions d is zero; that is, no constraints on the probabilities $\mathsf{P}_\theta(d = i)$ are imposed if $\theta \in \mathsf{I}_{\text{in}}$. The introduction of an indifference zone is typically motivated by the recognition that in many applications, the correct action is not crucial and often not even feasible when the hypotheses are very close. However, in principle, I_{in} may be an empty set.

We aim to find a sequential test $\delta = (T, d)$ that minimizes the expected sample size $\mathsf{E}_\theta[T]$ uniformly for all $\theta \in \Theta$ in the class of tests $\mathbb{C}(\alpha_0, \alpha_1)$ in which the maximal error probabilities $\sup_{\theta \in \Theta_i} \mathsf{P}_\theta(d \neq i)$ are upper-bounded by the given values:

$$\mathbb{C}(\alpha_0, \alpha_1) = \left\{ \delta : \sup_{\theta \in \Theta_0} \mathsf{P}_\theta(d = 1) \leqslant \alpha_0 \text{ and } \sup_{\theta \in \Theta_1} \mathsf{P}_\theta(d = 0) \leqslant \alpha_1 \right\}. \tag{2.31}$$

Thus, we are interested in the frequentist problem of finding a test δ_{opt} such that

$$\inf_{\delta \in \mathbb{C}(\alpha_0, \alpha_1)} \mathsf{E}_\theta[T] = \mathsf{E}_\theta[T_{\text{opt}}] \quad \text{uniformly in } \theta \in \Theta. \tag{2.32}$$

Unfortunately, such a uniformly optimal solution does not exist, and one has to resort to finding asymptotic approximations for small error probabilities. In the frequentist setting, it is possible to find first-order asymptotically optimal tests that satisfy

$$\lim_{\alpha_{\max} \to 0} \frac{\inf_{\delta \in \mathbb{C}(\alpha_0, \alpha_1)} \mathsf{E}_\theta[T]}{\mathsf{E}_\theta[T]} = 1 \quad \text{for all } \theta \in \Theta. \tag{2.33}$$

In addition to the frequentist problems (2.32) and (2.33), it is of interest to consider a Bayesian approach putting an a priori distribution $W(\theta)$ on Θ with a cost c per observation and a loss function $L(\theta)$ at the point θ associated with accepting the incorrect hypothesis and find asymptotically optimal tests when the cost c is small. The Bayes average (integrated) risk of a sequential test $\delta = (T, d)$ is

$$\rho_c^W(\delta) = \int_{\theta \leqslant \theta_0} L(\theta) \mathsf{P}_\theta(d = 1) W(\mathrm{d}\theta) + \int_{\theta \geqslant \theta_1} L(\theta) \mathsf{P}_\theta(d = 0) W(\mathrm{d}\theta) + c \int_\Theta \mathsf{E}_\theta[T] W(\mathrm{d}\theta).$$

It turns out that in the Bayesian context, it is possible to find tests that are not only asymptotically (as $c \to 0$) first-order optimal, $\inf_\delta \rho_c^W(\delta) = \rho_c^W(\delta)(1 + o(1))$, but also second-order optimal—that is, $\inf_\delta \rho_c^W(\delta) = \rho_c^W(\delta) + O(c)$—and even third-order optimal; that is, $\inf_\delta \rho_c^W(\delta) = \rho_c^W(\delta) + o(c)$.

In the case of the one-parameter exponential family (2.30), using optimal stopping theory, it can be shown that the optimal Bayesian test $\delta_{\text{opt}} = (T_{\text{opt}}, d_{\text{opt}})$ is

$$T_{\text{opt}} = \inf\{n \geqslant 1 : (S_n, n) \in \mathscr{B}_c\}, \quad d_{\text{opt}} = j \text{ if } (S_n, n) \in \mathscr{B}_c^j, \quad j = 0, 1,$$

where $S_n = X_1 + \cdots + X_n$, and $\mathscr{B}_c = \mathscr{B}_c^0 \cup \mathscr{B}_c^1$ is a set that can be found numerically.

Schwarz (1962) derived the test $\delta^\star(\hat{\theta})$ with $\hat{\theta} = \{\hat{\theta}_n\}$ being the maximum likelihood estimator (MLE) of θ as an asymptotic solution as $c \to 0$ to the Bayesian problem with the $0 - 1$ loss function. Specifically, the a posteriori risk of stopping is

$$R_n^{\text{st}}(S_n) = \min_{i = 0, 1} \left\{ \frac{\int_{\Theta_i} \exp\{\theta S_n - n b(\theta)\} \ W(\mathrm{d}\theta)}{\int_\Theta \exp\{\theta S_n - n b(\theta)\} \ W(\mathrm{d}\theta)} \right\}, \tag{2.34}$$

where $\Theta_0 = \{\theta \leqslant \theta_0\}$, $\Theta_1 = \{\theta \geqslant \theta_1\}$. Schwarz showed that $\mathscr{B}_c / |\log c| \to \mathscr{B}_0$ as $c \to 0$ and proposed a simple procedure: continue sampling until $R_n^{\text{st}}(S_n)$ is less than c and upon stopping accept the hypothesis for which the minimum is attained in equation (2.34). Denote this procedure by $\tilde{\delta}(c) = (\tilde{T}(c), \tilde{d}(c))$. Applying Laplace's asymptotic integration method to evaluate the integrals in equation (2.34) leads to the likelihood ratio test where the true parameter is replaced by the MLE $\hat{\theta}_n$. This approximation prescribes stopping sampling at the time $\hat{T}(\hat{\theta}) = \min(\hat{T}_0(\hat{\theta}), \hat{T}_1(\hat{\theta}))$, where

$$\hat{T}_i(\hat{\theta}) = \inf\{n : \sup_{\theta \in \Theta}[\theta S_n - n b(\theta)] - [\theta_i S_n - n b(\theta_i)] \geqslant |\log c|\}. \tag{2.35}$$

The terminal decision rule $\hat{d}(\hat{\theta})$ of the test $\hat{\delta}(\hat{\theta}) = (\hat{T}(\hat{\theta}), \hat{d}(\hat{\theta}))$ accepts $\mathcal{H}_0$ if $\hat{\theta}_{\hat{T}} < \theta^*$, where θ^* is such that $I(\theta^*, \theta_0) = I(\theta^*, \theta_1)$. Note also that

$$\hat{T} = \inf\left\{n \geqslant 1 : n \max\left[I(\hat{\theta}_n, \theta_0), I(\hat{\theta}_n, \theta_1)\right] \geqslant |\log c|\right\}. \tag{2.36}$$

The tests that use the MLEs of unknown parameters are usually referred to as the GSLRT.

Wong (1968) showed that the GSLRT $\hat{\delta}$ is first-order asymptotically Bayes as $c \to 0$:

$$\rho_c^W(\hat{\delta}) \sim \inf_\delta \rho_c^W(\delta) \sim c |\log c| \int_\Theta \frac{W(\mathrm{d}\theta)}{I_{\max}(\theta)}, \quad \mathsf{E}_\theta[\hat{T}] \sim \frac{|\log c|}{I_{\max}(\theta)} \quad \text{for every } \theta \in \Theta,$$

where $I_{\max}(\theta) = \max\{I(\theta, \theta_0), I(\theta, \theta_1)\}$.

Kiefer and Sacks (1963) showed that the procedure $\tilde{\delta}(c) = (\tilde{T}(c), \tilde{d}(c))$ with the stopping time $\tilde{T}(c) = \inf\{n \geqslant 1 : R_n^{\mathrm{st}}(S_n) \leqslant c\}$, proposed by Schwarz (1962), is also first-order asymptotically Bayes. In other words, for any prior distribution W, $\rho_c^W(\tilde{\delta}(c))$ behaves asymptotically like $\inf_\delta \rho_c^W(\delta)$ as $c \to 0$. Lorden (1967) refined this result by introducing the stopping region as the first n such that $R_n^{\mathrm{st}}(S_n) \leqslant Qc$, where Q is a positive constant, and demonstrated that it can be made second-order asymptotically optimal; that is, $\inf_\delta \rho_c^W(\delta) = \rho_c^W(\tilde{\delta}(Qc)) + O(c)$ as $c \to 0$, and $\inf_\delta \rho_c^W(\delta) = O(c|\log c|)$. It is noteworthy that the problem addressed by Lorden (1967) is more general than what we are discussing here because it encompasses general i.i.d. models, not limited to exponential families, and multiple-decision cases. Additionally, see Lorden (1972) for multiple hypotheses in one-parameter exponential families.

A significant advancement in Bayesian theory for testing separated hypotheses about the parameter of the one-parameter exponential family given by the equation (2.30) was made by Lorden (1977a; an unpublished manuscript). In this work, Lorden demonstrated that the family of GSLRTs can be devised to ensure third-order asymptotic optimality. This implies that they achieve the Bayes risk to within $o(c)$ as $c \to 0$.

Lorden provided sufficient conditions for families of tests to be third-order asymptotically Bayes and presented examples of such procedures based not only on the GLR approach but also on mixtures of likelihood ratios. Furthermore, the error probabilities of the GSLRTs were evaluated asymptotically as a consequence of a general theorem on boundary-crossing probabilities.

Due to the significance of this work, let us delve into a more detailed overview of Lorden's theory. It is worth noting that the paper by Lorden (1977a) extends the results obtained by Lorden (1977b) for multiple discrete cases, which we discussed in Subsection 2.1.2, to the continuous parameter case.

The hypotheses to be tested are $\mathcal{H}_0 : \underline{\theta} \leqslant \theta \leqslant \theta_0$ and $\mathcal{H}_1 : \bar{\theta} \geqslant \theta \geqslant \theta_1$, where $\underline{\theta}$ and $\bar{\theta}$ are interior points of the natural parameter space Θ. Let $\hat{\theta}_n \in [\underline{\theta}, \bar{\theta}]$ be the MLE that maximizes the likelihood over θ in $[\underline{\theta}, \bar{\theta}]$. Lorden's GSLRT stops at $\hat{T}$, which is the minimum of the Markov times $\hat{T}_0, \hat{T}_1$ defined as

$$\hat{T}_0(\hat{\theta}) = \inf\left\{ n \geqslant 1 : \sum_{k=1}^{n} \log\left[\frac{f_{\hat{\theta}_n}(X_k)}{f_{\theta_0}(X_k)} \; h_0(\hat{\theta}_n)\right] \geqslant a \text{ and } \hat{\theta}_n \geqslant \theta^* \right\},$$

$$\hat{T}_1(\hat{\theta}) = \inf\left\{ n \geqslant 1 : \sum_{k=1}^{n} \log\left[\frac{f_{\hat{\theta}_n}(X_k)}{f_{\theta_1}(X_k)} \; h_1(\hat{\theta}_n)\right] \geqslant a \text{ and } \hat{\theta}_n \leqslant \theta^* \right\},$$

(2.37)

where a is a threshold, θ^* satisfies $I(\theta^*, \theta_0) = I(\theta^*, \theta_1)$, and h_0, h_1 are positive continuous functions on $[\theta^*, \bar{\theta}], [\underline{\theta}, \theta^*]$, respectively. The hypothesis $\mathcal{H}_i$ is rejected when $\hat{T} = \hat{T}_i$. To summarize, Lorden's family of GSPRTs is defined as

$$\hat{T}(\hat{\theta}) = \min\left\{\hat{T}_0(\hat{\theta}), \hat{T}_1(\hat{\theta})\right\}, \quad \hat{d} = \begin{cases} 0 & \text{if } \hat{T}(\hat{\theta}) = \hat{T}_1(\hat{\theta}) \\ 1 & \text{if } \hat{T}(\hat{\theta}) = \hat{T}_0(\hat{\theta}) \end{cases},$$

(2.38)

with the $\hat{T}_i(\hat{\theta})$'s in equation (2.37).

Denote by

$$\lambda_n(\theta, \theta_i) = \sum_{k=1}^{n} \log\left[\frac{f_\theta(X_k)}{f_{\theta_i}(X_k)}\right] = (\theta - \theta_i)S_n - [b(\theta) - b(\theta_i)]n$$

the LLR between points θ and θ_i.

Lorden assumed that the prior distribution $W(\theta)$ has a continuous density $w(\theta)$ positive on $[\underline{\theta}, \bar{\theta}]$ and that the loss $L(\theta)$ equals zero in the indifference zone (θ_0, θ_1) and is continuous and positive elsewhere and bounded away from 0 on $[\underline{\theta}, \theta_0] \cup [\theta_1, \bar{\theta}]$. The main results in Lorden (1977a, theorem 1) can be briefly outlined as follows:

i. Under these assumptions, the family of GSLRTs defined by equations (2.37) and (2.38) with $a = |\log c| - \frac{1}{2}\log|\log c|$ is second-order asymptotically optimal; that is,

$$\rho_c^w(\hat{\delta}) = \inf_\delta \rho_c^w(\delta) + O(c) \quad \text{as } c \to 0,$$

where $\rho_c^w(\delta)$ is the average risk of the test $\delta = (T, d)$:

$$\rho_c^w(\delta) = \int_{\underline{\theta}}^{\theta_0} L(\theta)\mathsf{P}_\theta(d = 1)w(\theta)\mathrm{d}\theta + \int_{\theta_1}^{\bar{\theta}} L(\theta)\mathsf{P}_\theta(d = 0)w(\theta)\mathrm{d}\theta + c\int_{\underline{\theta}}^{\bar{\theta}} \mathsf{E}_\theta[T]w(\theta)\mathrm{d}\theta.$$

ii. This result can be improved from $O(c)$ to $o(c)$; that is, to the third order,

$$\rho_c^w(\hat{\delta}) = \inf_\delta \rho_c^w(\delta) + o(c) \quad \text{as } c \to 0,$$

making the right choice of the functions h_0 and h_1 by setting

$$h_i(\theta) = \sqrt{\frac{2\pi}{I^3(\theta, \theta_i)\ddot{b}(\theta)}} \; \frac{w(\theta)|\dot{b}(\theta) - \dot{b}(\theta_i)|}{w(\theta_i)L(\theta_i)\zeta(\theta, \theta_i)}, \quad i = 0, 1,$$

where $\zeta(\theta, \theta_i) = \mathcal{L}(\theta, \theta_i)/I(\theta, \theta_i)$ is a correction for the overshoot over the boundary, the factor that is the subject of renewal theory. Specifically,

$$\zeta(\theta, \theta_i) = \lim_{a \to \infty} \mathsf{E}_\theta \exp\left\{-\left[\lambda_{\tau_a}(\theta, \theta_i) - a\right]\right\}, \quad \tau_a = \inf\{n : \lambda_n(\theta, \theta_i) \geqslant a\}, \qquad (2.39)$$

where in the non-arithmetic case $\zeta(\theta, \theta_i)$ can be computed as

$$\zeta(\theta, \theta_i) = \frac{1}{I(\theta, \theta_i)} \exp\left\{-\sum_{n=1}^{\infty} \frac{1}{n} \left[\mathsf{P}_\theta(\lambda_n(\theta, \theta_i) \leqslant 0) + \mathsf{P}_{\theta_i}(\lambda_n(\theta, \theta_i) > 0)\right]\right\}. \qquad (2.40)$$

Because the Bayes average risk $\inf_\delta \rho_c^w(\delta)$ is of order $c|\log c|$, this implies that the asymptotic relative efficiency $\mathcal{E}_c = [\rho_c^w(\hat{\delta}) - \inf_\delta \rho_c^w(\delta)]/\rho_c^w(\hat{\delta})$ of Lorden's test is of order $1 - o(1/|\log c|)$ as $c \to 0$.

Note the crucial difference between Schwarz's GSLRT (equation 2.35) and Lorden's GSLRT (equation 2.38). In the Schwarz test, $h_i \equiv 1$ and the threshold is set as $a = |\log c|$. However, in the Lorden test, two innovations emerge. Firstly, the threshold is reduced by $\frac{1}{2}\log|\log c|$ and, secondly, adaptive weights $h_i(\hat{\theta}_n)$ are incorporated into the GLR statistic. Because the stopping times $\hat{T}_i$ can be written as

$$\begin{aligned}
\hat{T}_0(\hat{\theta}) &= \inf\left\{n \geqslant 1 : \lambda_n(\hat{\theta}_n, \theta_0) \geqslant a - \log h_0(\hat{\theta}_n) \text{ and } \hat{\theta}_n \geqslant \theta^*\right\}, \\
\hat{T}_1(\hat{\theta}) &= \inf\left\{n \geqslant 1 : \lambda_n(\hat{\theta}_n, \theta_1) \geqslant a - \log h_1(\hat{\theta}_n) \text{ and } \hat{\theta}_n \leqslant \theta^*\right\},
\end{aligned} \qquad (2.41)$$

Lorden's GSLRT can alternatively be perceived as the GSLRT with curved adaptive boundaries

$$a_i(\hat{\theta}_n) = |\log c| - \frac{1}{2}\log|\log c| - \log h_i(\hat{\theta}_n), \quad i = 0, 1,$$

which depend on the behavior of the MLE $\hat{\theta}_n$. These two innovations render this modification of the GLR test nearly optimal.

Given the complexity of Lorden's formal mathematical proof, we offer a heuristic sketch that captures the main ideas of the approach. The Bayesian perspective naturally guides us toward the mixture LR statistics

$$\bar{\Lambda}_n^i = \frac{\int_{\underline{\theta}}^{\bar{\theta}} e^{\theta S_n - nb(\theta)} w(\theta) \mathrm{d}\theta}{\int_{\Theta_i} L(\theta) e^{\theta S_n - nb(\theta)} w(\theta) \mathrm{d}\theta}, \quad i = 0, 1,$$

where $\Theta_0 = [\underline{\theta}, \theta_0]$, $\Theta_1 = [\theta_1, \bar{\theta}]$, and $L(\theta) = 1$ for the simple $0 - 1$ loss function. Indeed, the a posteriori stopping risk is given by

$$R_n^{\mathrm{st}}(S_n) = \min_{i=0,1} \left\{\frac{\int_{\Theta_i} L(\theta) e^{\theta S_n - nb(\theta)} w(\theta) \mathrm{d}\theta}{\int_{\underline{\theta}}^{\bar{\theta}} e^{\theta S_n - nb(\theta)} w(\theta) \mathrm{d}\theta}\right\}. \qquad (2.42)$$

A candidate for the approximate optimum is the procedure that stops as soon as $R_n^{\mathrm{st}}(S_n) \leqslant A_c$ for some $A_c \approx c$. This is equivalent to stopping as soon as $\max_{i=0,1} \bar{\Lambda}_n^i \geqslant 1/A_c$. The GLR statistics are approximated as

$$\hat{\Lambda}_n^i = \frac{\max_{\theta \in [\underline{\theta}, \bar{\theta}]} e^{\theta S_n - nb(\theta)}}{\max_{\theta \in \Theta_i} e^{\theta S_n - nb(\theta)}} \approx \frac{\max_{\theta \in [\underline{\theta}, \bar{\theta}]} e^{\theta S_n - nb(\theta)}}{e^{\theta_i S_n - nb(\theta_i)}}, \quad i = 0, 1,$$

and the stopping posterior risk (equation 2.42) is approximated as

$$R_n^{\text{st}}(S_n) \approx \min_{i=0,1} \frac{w(\theta_i) L(\theta_i) \left[\ddot{b}(\hat{\theta}_n)/2\pi n\right]^{1/2}}{w(\hat{\theta}_n) |\dot{b}(\theta_i) - \dot{b}(\hat{\theta}_n)|} \, e^{-\lambda_n(\hat{\theta}_n, \theta_i)}, \tag{2.43}$$

where $i = 0$ if $\hat{\theta}_n \leqslant \theta^*$ and $i = 1$ otherwise. These approximations stem from Laplace's method for asymptotic integral expansions and its variations.

Subsequently, Lorden demonstrated the existence of $Q > 1$ such that if the stopping risk exceeds Qc, then the continuation risk becomes smaller than the stopping risk. Therefore, it is approximately optimal to stop at the first instance such that R_n^{st} falls below Qc. This finding, coupled with the approximation given by equation (2.43), results in $T_{\text{opt}} \approx \min(\tau_0, \tau_1)$, where

$$\tau_i = \inf\left\{n : e^{-\lambda_n(\hat{\theta}_n, \theta_i)}/\tilde{h}_i(\hat{\theta}_n) n^{1/2} \leqslant Qc\right\} = \inf\left\{n : \lambda_n(\hat{\theta}_n, \theta_i) \geqslant -\log\left[n^{1/2} \tilde{h}_i(\hat{\theta}_n) Qc\right]\right\}$$

with $\tilde{h}_i(\hat{\theta}_n)$ given by

$$\tilde{h}_i(\hat{\theta}_n) = \sqrt{\frac{2\pi}{\ddot{b}(\hat{\theta}_n)}} \frac{w(\theta) |\dot{b}(\hat{\theta}_n) - \dot{b}(\theta_i)|}{w(\theta_i) L(\theta_i)}.$$

For small c, the expectation $\mathsf{E}_\theta[\tau_i]$ is of order $|\log c|$, so $n^{1/2}$ can be replaced by $|\log c|^{1/2}$, which yields

$$\tau_i \approx \hat{T}_i = \inf\left\{n : \lambda_n(\hat{\theta}_n, \theta_i) \geqslant -\log\left[c|\log c|^{1/2} Q\tilde{h}_i(\hat{\theta}_n)\right]\right\}.$$

Note that these stopping times look exactly like the ones defined in equation (2.41) with the stopping boundaries

$$a_i(\hat{\theta}_n) = |\log c| - \frac{1}{2}\log|\log c| - \log\left[Q\tilde{h}_i(\hat{\theta}_n)\right], \quad i = 0, 1.$$

The test based on these stopping times is already optimal to the second order. However, to achieve third-order optimality, one must carefully choose the constant Q to address the overshoots. Specifically, leveraging this result, Lorden demonstrated that the risks of an optimal rule and of the GSLRT are both linked to the risks of the family of one-sided tests $\tau_a(\theta) = \inf\{n : \lambda_n(\theta, \theta_i) \geqslant a\}$, which are strictly optimal in the problem

$$\rho(\theta, \nu) = \inf_T \left\{\mathsf{E}_\theta T + \nu \mathsf{P}_{\theta_i}(T < \infty)\right\} = \inf_T \mathsf{E}_\theta\left\{T + \nu \prod_{n=1}^T \frac{p_{\theta_i}(X_n)}{p_\theta(X_n)}\right\}.$$

If we set $a = \log[\nu \mathcal{L}(\theta, \theta_i)]$, then by taking $Q = 1/\mathcal{L}(\theta, \theta_i)$ the resulting test will be nearly optimal to within $o(c)$. Because θ is unknown, we need to replace it with the estimate $\hat{\theta}_n$ to obtain

$$a_i(\hat{\theta}_n) = |\log c| - \frac{1}{2}\log|\log c| - \log\Big[\tilde{h}_i(\hat{\theta}_n)/\mathcal{L}(\hat{\theta}_n, \theta_i)\Big]$$

$$= |\log c| - \frac{1}{2}\log|\log c| - \log\Big[h_i(\hat{\theta}_n)\Big].$$

It is intriguing to compare Lorden's approach with the Kiefer-Sacks test that stops the first time R_n^{st} becomes smaller than c. Lorden's approach allows us to show that the test with the stopping time

$$\hat{T} = \inf\Big\{n : R_n^{\mathrm{st}}(S_n) \leqslant c/\mathcal{L}(\hat{\theta}_n)\Big\},$$

where $\mathcal{L}(\hat{\theta}_n) = \mathcal{L}(\hat{\theta}_n, \theta_1)$ if $\hat{\theta}_n < \theta^*$ and $\mathcal{L}(\hat{\theta}_n) = \mathcal{L}(\hat{\theta}_n, \theta_0)$ otherwise, is nearly optimal to within $o(c)$. It is worth recalling that the factor $\zeta(\theta, \theta_i) = I(\theta, \theta_i)^{-1}\mathcal{L}(\theta, \theta_i)$ provides a necessary correction for the excess over the thresholds at stopping; see equation (2.39). This offers a significant enhancement over the Kiefer-Sacks test, which disregards the overshoots. Notably, this improvement is not limited to testing close hypotheses when $\mathcal{L}(\theta, \theta_i) \ll 1$. Even in cases where the parameter values are well separated, this correction could be crucial. For instance, in the binomial case with the success probabilities $\theta_1 = 0.6$ and $\theta_0 = 0.4$, we have $\mathcal{L}(\theta_1, \theta_0) \approx 1/15$, so Lorden's test will terminate much earlier.

Certainly, it is important to note that implementing Lorden's fully optimized GSLRT may encounter difficulties. This is primarily because computing the numbers $\zeta(\theta, \theta_i)$ analytically is often not feasible, except for specific models such as the exponential. For instance, when testing the mean in the Gaussian case, these numbers can only be computed numerically. Though Siegmund's (1985) corrected Brownian motion approximations can be utilized, they are sufficiently accurate only when the difference between θ and θ_i is relatively small. Hence, for practical purposes, only partially optimized solutions, which provide $O(c)$-optimality, are typically feasible. A workaround involves discretizing the parameter space.

Let $\hat{\alpha}_0(\theta) = \mathsf{P}_\theta(\hat{d} = 1)$, $\theta \in \Theta_0 = [\underline{\theta}, \theta_0]$ and $\hat{\alpha}_1(\theta) = \mathsf{P}_\theta(\hat{d} = 0)$, $\theta \in \Theta_1 = [\theta_1, \bar{\theta}]$ denote the error probabilities of the GSLRT $\hat{\delta}_a$. Note that due to the monotonicity of $\hat{\alpha}_i(\theta)$, $\sup_{\theta \in \Theta_i} \hat{\alpha}_i(\theta) = \hat{\alpha}_i(\theta_i)$. In addition to the Bayesian third-order optimality property, Lorden established asymptotic approximations to the error probabilities of the GSLRT. Specifically, by theorem 2 of Lorden (1977a),

$$\hat{\alpha}_i(\theta_i) = \sqrt{a}e^{-a}C_i(\theta_i)(1 + o(1)), \quad i = 0, 1 \quad \text{as } a \to \infty, \tag{2.44}$$

where

$$C_0(\theta_0) = \int_{\theta^*}^{\bar{\theta}} \zeta(\theta, \theta_0)h_0(\theta)\sqrt{\frac{\ddot{b}(\theta)}{2\pi I(\theta, \theta_0)}}\ \mathrm{d}\theta,$$

$$C_1(\theta_1) = \int_{\underline{\theta}}^{\theta^*} \zeta(\theta, \theta_1)h_1(\theta)\sqrt{\frac{\ddot{b}(\theta)}{2\pi I(\theta, \theta_1)}}\ \mathrm{d}\theta$$

and where $\zeta(\theta, \theta_i)$, $i = 0, 1$, are defined in equations (2.39) and (2.40). These approximations hold significance for frequentist problems, which are typically of primary interest in most applications. Though there are no strict upper bounds on the error

probabilities, leading to no specific prescription on how to embed the GSLRT into class $\mathbb{C}(\alpha_0, \alpha_1)$, the asymptotic approximations (equation 2.44) enable us to select thresholds a_i in the stopping times $\hat{T}_i$ so that $\hat{\alpha}_i(\theta_i) \approx \alpha_i$, $i = 0, 1$, at least for sufficiently small α_i. Note that in this latter case, the threshold a in equation (2.41) should be replaced with a_i, the roots of the transcendental equations

$$a_i - \frac{1}{2}\log a_i = \log[C_i(\theta_i)/\alpha_i], \quad i = 0, 1.$$

With this choice, the GSLRT is asymptotically uniformly first-order optimal with respect to the expected sample size; that is,

$$\inf_{\delta \in \mathbb{C}(\alpha_0, \alpha_1)} \mathsf{E}_\theta[T] = \mathsf{E}_\theta[\hat{T}](1 + o(1)) \quad \text{as } \alpha_{\max} \to 0 \quad \text{for all } \theta \in [\underline{\theta}, \bar{\theta}],$$

where the $o(1)$ term is of order $O(\log|\log \alpha_{\max}|/|\log \alpha_{\max}|)$. It is noteworthy that this result holds true not only in the asymptotically symmetric case where $\log \alpha_0 \sim \log \alpha_1$ and $a_0 \sim a_1$ as $\alpha_{\max} \to 0$ but also in the asymmetric case where a_0 and a_1 diverge with different rates, as long as $a_1 e^{-a_0} \to 0$.

Note that the Schwarz-Lorden asymptotic theory operates under the assumption of a fixed indifference zone that does not permit local alternatives, meaning that θ_1 cannot approach θ_0 as $c \to 0$. In simpler terms, this theory is confined to scenarios where the width of the indifference zone $\theta_1 - \theta_0$ is considerably larger than $c^{1/2}$.

We conclude this section by mentioning Lorden's (1973) paper on the properties of the one-sided (open-ended) GSLRTs for the one-parameter exponential family. These tests reject a null hypothesis $\theta = \theta_0$ in favor of $\theta > \theta_0$ within the class of stopping times satisfying $\mathsf{P}_{\theta_0}(T < \infty) \leqslant \alpha$ for a prescribed $0 < \alpha < 1/3$.

2.5. Optimal Multistage Testing

What is the fewest number of stages for which a multistage hypothesis test can be asymptotically equivalent to an optimal fully sequential test? Lorden (1983) took up this question and reached the definitive answer of needing three stages in general, except in a special symmetric situation in which two stages are possible, described in the next section. Here, "needing three stages" means allowing the *possibility* of three stages, although Lorden's optimal procedures can (and do, with probability approaching 1) terminate earlier; see Section 2.5.3. Lorden (1983) showed this first in the simple vs. simple testing setup and then for testing separated composite hypotheses in an exponential family. In this area again, Lorden's work was groundbreaking and formed the foundation for later, more general theoretical investigations in optimal multistage testing (e.g., Bartroff 2006a, 2006b, 2007; Xing and Fellouris 2023) and in applications to clinical trial designs where the problem is sometimes known as "sample size adjustment" or "re-estimation" (e.g., Bartroff and Lai 2008a, 2008b; Bartroff, Lai, and Shih 2013). In this literature especially, multistage procedures are often referred to as group sequential. Throughout this section, i.i.d. observations are assumed.

2.5.1. Simple vs. Simple Testing: Multistage Competitors of the SPRT

Beginning with the simple vs. simple testing setup of Section 2.1.1 and adopting the notation there, some of Lorden's main ideas can be seen by first considering the symmetric case where the error probabilities $\alpha_0, \alpha_1 \to 0$ in such a way that

$$\frac{\log \alpha_1^{-1}}{I_0} \sim \frac{\log \alpha_0^{-1}}{I_1}. \tag{2.45}$$

Letting λ_n be the LLR statistic in equation (2.3) and $t \to \infty$ an argument parameterizing $\alpha_0, \alpha_1 \to 0$, Lorden (1983) began by arguing that there is a sample size $n = n(t) \geqslant t$ such that $n = t + o(t)$,

$$\mathsf{P}_0(-\lambda_n < tI_0) \to 0, \quad \text{and} \quad \mathsf{P}_1(\lambda_n < tI_1) \to 0. \tag{2.46}$$

More explicitly, this is achievable by taking $n = t + \delta_t$ with

$$\sqrt{t} \ll \delta_t \ll t \tag{2.47}$$

because, assuming finite second moments $\mathsf{E}_i[\lambda_1^2] < \infty$, Chebyshev's inequality gives

$$\mathsf{P}_1(\lambda_n < tI_1) \leqslant \frac{\mathsf{var}_1(\lambda_1)/n}{I_1^2(1 - t/n)^2}$$

which, ignoring constants, under equation (2.47) is

$$\frac{1/n}{(1 - t/n)^2} = \frac{n}{\delta_t^2} = \frac{t + \delta_t}{\delta_t^2} = o(1) \quad \text{as } t \to \infty.$$

A similar argument shows that the other probability in equation (2.46) approaches 0 as well.

In this symmetric situation, an optimal two-stage competitor to the SPRT can be described in terms of $n(t)$, which is the size of the first stage with t taken to be the larger of the two sides of equation (2.45). Note that, for either $i = 0$ or 1, we have $tI_i \sim \log \alpha_{1-i}^{-1}$ so that t is asymptotically the same as the expected stopping time of the SPRT (under either hypothesis) and the first stage $n(t)$ is of the same order but slightly larger. The procedure stops after the first stage if

$$\lambda_{n(t)} \notin (\log \alpha_1, \log \alpha_0^{-1}), \tag{2.48}$$

making the appropriate terminal decision. Using equation (2.46), the probability under the null of terminating and making the correct terminal decision after this first stage is

$$\mathsf{P}_0(\lambda_{n(t)} \leqslant \log \alpha_1) = \mathsf{P}_0(-\lambda_{n(t)} \geqslant \log \alpha_1^{-1}) \geqslant \mathsf{P}_0(-\lambda_{n(t)} \geqslant tI_0) \to 1, \tag{2.49}$$

with a similar argument showing that

$$\mathsf{P}_1(\lambda_{n(t)} \geqslant \log \alpha_0^{-1}) \to 1. \tag{2.50}$$

Otherwise, the test continues to a total sample size n_2, which is that of the fixed-sample-size test with error probabilities α_0, α_1 and uses that terminal decision rule. This can be accomplished in at most $n_2 \leqslant Ct \leqslant Cn(t)$ total observations for some constant C. Thus, under the null, the total expected sample size is at most

$$n(t) + Cn(t)\mathsf{P}_0(\lambda_{n(t)} > \log \alpha_1) = n(t)[1 + o(1)] \sim t \sim \frac{\log \alpha_1^{-1}}{I_0},$$

and is of the same order $I_1^{-1} \log \alpha_0^{-1}$ under the alternative by a similar argument. By definition of the two stages, the procedure has type I error probability at most $2\alpha_0$ and type II error probability at most $2\alpha_1$, so repeating the construction with $\alpha_i/2$ replacing α_i ($i = 0$, 1) controls the error probabilities at the nominal levels and does not affect the asymptotic estimates above. Thus, this two-stage procedure is asymptotically as efficient as the SPRT in this symmetric case.

If the asymptotic equivalence equation (2.45) does not hold but we assume that

$$\frac{\log \alpha_0^{-1}}{\log \alpha_1^{-1}} \quad \text{is bounded away from 0 and } \infty, \qquad (2.51)$$

Lorden showed that no two-stage test can be asymptotically optimal, itself a nontrivial result that we discuss in the next section. For this case Lorden gives a three-stage procedure that is a slight modification of the one above. Letting t_1 and t_2 be the left- and right-hand sides of equation (2.45), respectively, the first stage of the procedure is of size $\min\{n(t_1), n(t_2)\}$, and the second stage (if needed) brings the total sample size to $\max\{n(t_1), n(t_2)\}$, both using the stopping rule equation (2.48) and corresponding decision rule. If not stopped by the second stage, a third stage brings the total sample size to that of the fixed sample size with error probabilities α_i, which is $\leqslant C\max\{n(t_1), n(t_2)\}$ as above, and uses that terminal decision rule. Because $n(t_{i+1}) \sim t_{i+1} \sim \log \alpha_{1-i}^{-1}/I_i$ for both $i = 0$, 1, and equations (2.49) and (2.50) hold for $n(t_1)$ and $n(t_2)$, respectively, the expected sample size of this three-stage procedure is asymptotically equal to the corresponding side of equation (2.45) and is thus minimized under both the null and alternative.

2.5.2. The Necessity of Three Stages

Continuing with the simple vs. simple testing setup of the previous section, Lorden's (1983, corollary 1) result mentioned above that, in the absence of symmetry equation (2.45), three stages are necessary for asymptotic optimality is far from obvious because it may seem that the first two stages of the three-stage procedure defined above would suffice. That is, why is it that a first stage of $\min\{n(t_1), n(t_2)\}$ and (if needed) a second stage giving total sample size $\max\{n(t_1), n(t_2)\}$ would not be optimal? One clue may be that, if that were true, then the same reasoning would seem to imply that a single-stage test could be optimal under symmetry equation (2.45), which is known to not hold. More generally, Lorden provided the following general result about asymptotically optimal k-stage ($k \geqslant 2$) tests: that their expected sample size after $k-1$ stages must be asymptotically the same as after k stages. In other words, the final stage of an asymptotically optimal multistage test is asymptotically negligible in size but *necessary*. In what follows, let $I(f, g)$ denote the information number for arbitrary densities f, g.

Theorem 2.5 (Lorden 1983, theorem 3). *For testing f_0 vs. f_1 in the setup of Section 2.1.1, let N denote the sample size of a k-stage ($k \geqslant 2$) test with error probabilities α_0 and α_1, and let M be the total sample size of this test after $k{-}1$ stages. If N is asymptotically optimal as $\alpha_0, \alpha_1 \rightarrow 0$ and g is a density distinct from f_0 such that*

$$\frac{\log \alpha_1^{-1}}{\log \alpha_0^{-1}} \geqslant Q > \frac{I(g, f_1)}{I(g, f_0)}$$

for some $Q > 0$ as $\alpha_0, \alpha_1 \to 0$, then

$$M \to \frac{\log \alpha_0^{-1}}{I(g, f_0)} \text{ in } g\text{-probability, and } \mathsf{E}_g[M] \sim \frac{\log \alpha_0^{-1}}{I(g, f_0)} \sim \mathsf{E}_g[N] \quad \text{as } \alpha_0 \to 0.$$

Lorden's proof of this theorem is technical and requires detailed upper bounds on the *conditional* error probabilities after the $(k - 1)$st stage; that is, the probabilities of test error given the first M observations. Roughly speaking, showing that these error probabilities are small shows that their corresponding sample size M must be large, so large in fact that it is asymptotically equivalent to its maximum value N.

Lorden (1983, corollary 1) then used Theorem 2.5 to show that there is an asymptotically optimal two-stage test *if and only if* the symmetry condition equation (2.45) holds, with the construction of the two-stage test above providing the "if" argument. For the converse, applying Theorem 2.5 with $g = f_1$ shows that the first stage of an optimal two-stage test must be asymptotic to $(\log \alpha_0^{-1})/I_1$. After reversing the roles of f_0 and f_1 in the theorem and applying it again with $g = f_0$, it also shows that the first stage must be asymptotic to $(\log \alpha_1^{-1})/I_0$, establishing symmetry equation (2.45).

2.5.3. Composite Hypotheses

For testing separated hypotheses $\theta \leqslant \theta_0$ vs. $\theta \geqslant \theta_1 > \theta_0$ about the one-dimensional parameter θ of an exponential family, Lorden (1983, section 3) constructed an asymptotically optimal three-stage test utilizing a description of the optimal stopping boundary related to Schwarz's (1962) study of Bayes asymptotic shapes for fully sequential tests, described in Section 2.4. Let $n(\theta)$ denote the expected sample size to Schwarz's boundary under θ. Lorden's test utilizes the "worst-case" competing parameter value $\theta^* \in (\theta_0, \theta_1)$ that maximizes the expected sample size $n(\theta^*) = \max_\theta n(\theta) \equiv n^\star$. The first stage size of Lorden's procedure is a fixed fraction of $n^\star$. If the procedure does not stop after the first stage, utilizing Schwarz's boundary, the second stage brings the total sample size to $\min\{n^\star, (1 + \varepsilon)n(\hat{\theta})\}$, where $\hat{\theta}$ is the MLE of θ from the first-stage data and $\varepsilon \searrow 0$ is a chosen sequence. Finally, if needed, the third stage brings the total sample size up to $n^\star$. Under equation (2.51), Lorden (1983, theorem 1) proved that this test asymptotically minimizes the expected sample size to first order, not just for θ in the hypotheses but uniformly in θ over any interval in the parameter space containing $[\theta_0, \theta_1]$. The first order term is of order $\log \alpha_i^{-1}$, as above, and the second order term is of order $O(((\log \alpha_i^{-1}) \log \log \alpha_i^{-1})^{1/2})$, $i = 0, 1$.

These results were extended to asymptotically optimal three-stage tests of multidimensional parameters in Bartroff (2006a) and Bartroff and Lai (2008a), and more general multidimensional composite hypotheses in Bartroff and Lai (2008b). On the other hand, Lorden's procedures were generalized to optimal k-stage tests, for arbitrary $k \geqslant 3$, in Bartroff (2006b, 2007).

Regarding the necessity of three stages in this composite hypothesis setting, Lorden (1983, corollary 2) proved that, under equation (2.51), three stages are necessary (and sufficient, by his own procedure) for asymptotic optimality at more than three values of θ, and so certainly for asymptotic optimality over an interval of θ values, as in Lorden's result. An interesting detail that shows this result to be best possible is that an optimal

two-stage test *can* be constructed at three values of θ if the special symmetry condition $I(\theta', \theta_0)I(\theta_0, \theta_1) = I(\theta', \theta_1)I(\theta_1, \theta_0)$ holds for some $\theta' \neq \theta_0, \theta_1$. Then a two-stage procedure similar to the one described in Section 2.5.1 that uses second-stage total sample size of $\log \alpha_0^{-1}/I(\theta', \theta_0)$ will be optimal at the three values $\theta = \theta'$, θ_0, and θ_1.

3. SEQUENTIAL CHANGE POINT DETECTION: LORDEN'S MINIMAX CHANGE DETECTION THEORY

In numerous practical applications, the observed process undergoes an abrupt change in statistical properties at an unknown point in time. Examples encompass aerospace navigation and flight systems integrity monitoring, cybersecurity, identification of terrorist activity, industrial monitoring, air pollution monitoring, radar, sonar, and electro-optics surveillance systems. Consequently, this problem has garnered interest from many practitioners for some time.

In classical quickest change point detection, the objective is to detect changes in the distribution as swiftly as possible, thereby minimizing the expected delay to detection assuming that the change is in effect.

More specifically, the change point problem posits that one obtains a series of observations $X_1, X_2, ...$, such that, for some value ν, $\nu \in \mathbb{Z}_+ = \{0, 1, 2, ...\}$ (the change point), $X_1, X_2, ..., X_\nu$ have one distribution and $X_{\nu+1}, X_{\nu+2}, ...$, have another distribution. The change point ν is unknown, and the sequence $\{X_n\}_{n \geqslant 1}$ is being monitored for detecting a change. A sequential detection procedure is a stopping time T with respect to the Xs, so that after observing $X_1, X_2, ..., X_T$ it is declared that a change is in effect. That is, T is an integer-valued random variable, such that the event $\{T = n\}$ belongs to the sigma-algebra $\mathscr{F}_n = \sigma(X_1, ..., X_n)$ generated by observations $X_1, ..., X_n$.

Historically, the field of change point detection began to take shape in the 1920s to 1930s, spurred by considerations in quality control. Shewhart's charts were particularly influential during this period (Shewhart 1931). However, optimal and nearly optimal sequential detection procedures did not come into prominence until much later, in the 1950s to 1970s, following the advent of sequential analysis (Wald 1947). The concepts initiated by Shewhart and Wald laid the foundation for extensive research into sequential change point detection.

The desire to detect the change quickly often leads to being "trigger-happy," which, on one hand, results in an unacceptably high false alarm rate; that is, terminating the process prematurely before a real change has occurred. On the other hand, attempting to avoid false alarms too strenuously causes a long delay between the true change point and its detection. Thus, the essence of the problem lies in achieving a trade-off between two conflicting performance measures: the loss associated with the delay in detecting a true change and that associated with raising a false alarm. An efficient detection procedure is expected to minimize the average loss associated with the detection delay, while subject to a constraint on the loss associated with false alarms, or vice versa.

Let $p_\nu(\mathbf{X}^n) = p(X_1, ..., X_n | \nu)$ denote the joint probability density of the sample $\mathbf{X}^n = (X_1, ..., X_n)$ when the change point ν is fixed ($0 \leqslant \nu < \infty$) and $p_\infty(\mathbf{X}^n)$ the joint density when $\nu = \infty$; that is, when there is never a change. Let $\mathsf{P}_\nu, \mathsf{P}_\infty$ and $\mathsf{E}_\nu, \mathsf{E}_\infty$ denote the

corresponding probability measures and expectations. Assume that the observations $\{X_n\}_{n \geqslant 1}$ are independent and such that $X_1, ..., X_\nu$ are each distributed according to a common (prechange) density $f_0(x)$, while $X_{\nu+1}, X_{\nu+2}, ...$, each follow a common (postchange) density $f_1(x)$. Hence, the model can be represented as

$$p_\nu(\mathbf{X}^n) = \begin{cases} \prod_{t=1}^{\nu} f_0(X_t) \times \prod_{t=\nu+1}^{n} f_1(X_) & \text{for } n \geqslant \nu + 1 \\ p_\infty(\mathbf{X}^n) = \prod_{t=1}^{n} f_0(X_t) & \text{for } 1 \leqslant n \leqslant \nu \end{cases}. \tag{3.1}$$

Note that we assume that X_ν is the last prechange observation, which is different from many publications (including Lorden's) where it is assumed that X_ν is the first postchange observation. The diagram below illustrates this case:

$$\underbrace{X_1, ..., X_\nu}_{\text{i.i.d.}, f_0}, \underbrace{X_{\nu+1}, X_{\nu+2}, ...}_{\text{i.i.d.}, f_1}.$$

Denote by $\mathcal{H}_\infty : \nu = \infty$ the hypothesis that the change never occurs and by $\mathcal{H}_\nu$ the hypothesis that the change occurs at time $0 \leqslant \nu < \infty$. Let $Z_t = \log[f_1(X_t)/f_0(X_t)]$ denote the LLR for the tth observation X_t.

We now introduce the cumulative sum (CUSUM) detection procedure, which was first proposed by Page (1954). The change point detection problem can be viewed as a problem of testing two hypotheses: $\mathcal{H}_\nu$ that the change occurs at a fixed point $0 \leqslant \nu < \infty$ against the alternative $\mathcal{H}_\infty : \nu = \infty$ that the change never occurs. The LLR between these hypotheses is $\lambda_n^\nu = \sum_{t=\nu+1}^{n} Z_t$ for $\nu < n$ and 0 for $\nu \geqslant n$. Because hypothesis $\mathcal{H}_\nu$ is composite, we may employ the GLR approach, maximizing the LLR λ_n^ν over ν, to obtain the log-GLR statistic:

$$W_n = \max_{\nu \geqslant 0} \sum_{t=\nu+1}^{n} Z_t, \tag{3.2}$$

which follows the recursion

$$W_n = (W_{n-1} + Z_n)^+, \quad n \geqslant 1, W_0 = 0. \tag{3.3}$$

This statistic is called the CUSUM statistic. Page's CUSUM procedure is the first time $n \geqslant 1$ such that the CUSUM statistic W_n exceeds a positive threshold a:

$$T_a = \inf\{n \geqslant 1 : W_n \geqslant a\}. \tag{3.4}$$

Page (1954) proposed measuring the risk due to a false alarm by the mean time to false alarm $\mathsf{E}_\infty[T]$ and the risk associated with a true change detection by the mean time to detection $\mathsf{E}_0[T]$ when the change occurs at the very beginning. These are commonly known as the average run length (ARL). Page also analyzed the CUSUM procedure defined by equations (3.2) to (3.4) using these operating characteristics.

Though the false alarm rate is reasonable to measure by the ARL to false alarm $\mathsf{ARLFA}(T) = \mathsf{E}_\infty[T]$, the risk due to a true change detection is better measured by the conditional expected delay to detection $\mathsf{E}_\nu[T - \nu | T > \nu]$ for any possible change point $\nu \in \mathbb{Z}_+$, rather than by the ARL to detection $\mathsf{E}_0[T]$. Ideally, a good detection procedure

should guarantee small values of the expected detection delay for all change points $\nu \in \mathbb{Z}_+$ when $\mathsf{ARLFA}(T)$ is set at a certain level. However, if the false alarm risk is measured in terms of the ARL to false alarm—that is, it is required that $\mathsf{ARLFA}(T) \geqslant \gamma$ for some $\gamma \geqslant 1$—then a procedure that minimizes the conditional expected delay to detection $\mathsf{E}_\nu[T - \nu | T > \nu]$ uniformly over all ν does not exist. For this reason, we must resort to different optimality criteria, such as Bayesian and minimax criteria.

The minimax approach posits that the change point is an unknown not necessarily random number. Even if it is random, its distribution is unknown.

Lorden (1971) was the first to address the minimax change detection problem and developed the first minimax theory. He proposed to measure the false alarm risk by the ARL to false alarm $\mathsf{ARLFA}(T) = \mathsf{E}_\infty[T]$; that is, to consider the class of change detection procedures $\mathbb{C}(\gamma) = \left\{ T : \mathsf{ARLFA}(T) \geqslant \gamma \right\}$ for some $\gamma \geqslant 1$ and the risk associated with detection delay by the worst-case expected detection delay

$$\mathsf{ESEDD}(T) = \sup_{0 \leqslant \nu < \infty} \left\{ \text{ess sup } \mathsf{E}_\nu \left[(\mathrm{T} - \nu)^+ | X_1, ..., X_\nu \right] \right\}. \tag{3.5}$$

In other words, the conditional expected detection delay is maximized over all possible trajectories $(X_1, ..., X_\nu)$ up to the change point and then over the change point ν.

Lorden's minimax criterion is

$$\inf_{T} \sup_{\nu \geqslant 0} \text{ess sup}_\omega \mathsf{E}_\nu[\mathrm{T}\text{-}\nu | \mathrm{T} > \nu, \mathscr{F}_\nu] \quad \text{subject to } \mathsf{ARLFA}(\mathrm{T}) \geqslant \gamma,$$

that is, Lorden's minimax optimization problem seeks to

$$\text{Find } T_{\text{opt}} \in \mathbb{C}(\gamma) \text{ such that } \mathsf{ESEDD}(T_{\text{opt}}) = \inf_{T \in \mathbb{C}(\gamma)} \mathsf{ESEDD}(T) \text{ for every } \gamma \geqslant 1. \tag{3.6}$$

Lorden (1971) demonstrated that Page's CUSUM procedure achieves first-order asymptotic minimax optimality as γ approaches infinity. This groundbreaking finding marked the initial optimality result in the minimax change detection problem. Given the significance of this outcome and the widespread adoption of Lorden's minimax criterion not only within statistical circles but also across various practical domains, we proceed to provide further elaboration.

To establish the asymptotic optimality of Page's CUSUM procedure, Lorden employed an intriguing method that permits the utilization of one-sided hypothesis tests to assess a collection of change detection procedures, among them Page's method. Let $\tau = \tau(\alpha)$ be a stopping time with respect to $X_1, X_2, ...,$ such that

$$\mathsf{P}_\infty(\tau < \infty) \leqslant \alpha, \tag{3.7}$$

where $\alpha \in (0, 1)$. For $k = 0, 1, 2, ...,$ define the stopping time τ_k obtained by applying τ to the sequence $X_{k+1}, X_{k+2}, ...,$ and let $\tau^* = \min_{k \geqslant 0}(\tau_k + k)$.

The subsequent theorem, resembling theorem 2 in Lorden (1971), empowers the construction of nearly optimal change detection procedures and facilitates the demonstration of the near-optimality of the CUSUM procedure. It is important to recall that P_∞ denotes the distribution characterized by the density $f_0(x)$, and P_0 corresponds to the distribution with density $f_1(x)$.

Theorem 3.1. *The random variable τ^* is a stopping time with respect to $X_1, X_2, ...,$ and if condition (3.7) is satisfied, then the following two inequalities hold:*

$$\mathsf{E}_\infty[\tau^*] \geqslant 1/\alpha \tag{3.8}$$

and

$$\mathsf{E}_0[\tau^*] \leqslant \mathsf{E}_0[\tau]. \tag{3.9}$$

The cumulative LLR for the sample $(X_{k+1}, ..., X_n)$ is $\lambda_n^k = \sum_{t=k+1}^n Z_t$. Let $\tau(\alpha) = \inf\{n \geqslant 1 : \lambda_n^0 \geqslant |\log \alpha|\}$ denote the stopping time of the one-sided SPRT for testing f_0 versus f_1 with threshold $|\log \alpha|$. Then $\mathsf{P}_\infty(\tau(\alpha) < \infty) \leqslant \alpha$, so condition (3.7) holds. If the Kullback-Leibler information number $I = \mathsf{E}_0[Z_1]$ is positive and finite, then it is well known that

$$\mathsf{E}_0[\tau(\alpha)] = \frac{|\log \alpha|}{I}(1 + o(1)) \quad \text{as } \alpha \to 0.$$

Next, note that the CUSUM statistic defined in equation (3.2) is the maximum of λ_n^k over $k \geqslant 0$, so the stopping time of the CUSUM procedure defined in equation (3.3) can obviously be written as $T_a = \min_{k \geqslant 0}\{\tau_k(\alpha) + k\} \equiv \tau^*$ for $a = a_\alpha = |\log \alpha|$, where

$$\tau_k(\alpha) = \inf\left\{n \geqslant 1 : \lambda_{k+n}^k \geqslant |\log \alpha|\right\}.$$

It follows from Theorem 3.1 that setting $\alpha = \gamma^{-1}$ gives $\mathsf{E}_\infty[T_{a_\gamma}] \geqslant \gamma$, so $T_{a_\gamma} \in \mathbb{C}(\gamma)$, and

$$\mathsf{ESEDD}(T_{a_\gamma}) \equiv \mathsf{E}_0\left[T_{a_\gamma}\right] = \frac{\log \gamma}{I}(1 + o(1)) \quad \text{as } \gamma \to \infty.$$

To complete the proof of the first-order asymptotic optimality of the CUSUM procedure with threshold $a = a_\gamma = \log \gamma$, it suffices to establish that this is the best one can do; that is, to prove the asymptotic lower bound

$$\inf_{T \in \mathbb{C}(\gamma)} \mathsf{ESEDD}(T) \geqslant \frac{\log \gamma}{I}(1 + o(1)) \quad \text{as } \gamma \to \infty, \tag{3.10}$$

which also yields

$$\inf_{T \in \mathbb{C}(\gamma)} \mathsf{ESEDD}(T) \sim \frac{\log \gamma}{I} \sim \mathsf{ESEDD}(T_{a_\gamma}) \quad \text{as } \gamma \to \infty.$$

Theorem 3 of Lorden (1971) establishes this fact using a rather sophisticated argument. Note, however, that Lai (1998) established the lower bound given by equation (3.10) in a general non-i.i.d. case, assuming that $n^{-1}\lambda_{\nu+n}^\nu$ converges to a positive and finite number I as $n \to \infty$ under a certain additional condition. In the i.i.d. case, by the Strong Law of Large Numbers (SLLN) $n^{-1}\lambda_{\nu+n}^\nu$ converges to the Kullback-Leibler information number I almost surely under P_ν. This implies that as $M \to \infty$ for all $\varepsilon > 0$,

$$\sup_{\nu \geqslant 0} \mathsf{P}_\nu\left\{\frac{1}{M} \max_{0 \leqslant n \leqslant M} \lambda_{\nu+n}^\nu \geqslant (1 + \varepsilon)I\right\} = \mathsf{P}_0\left\{\frac{1}{M} \max_{0 \leqslant n \leqslant M} \lambda_n^0 \geqslant (1 + \varepsilon)I\right\} \to 0. \tag{3.11}$$

Using equation (3.11), the lower bound equation (3.10) can be obtained from theorem 1 in Lai (1998).

To handle a composite parametric postchange hypothesis, which is typical in many applications, let $f_\theta(x)$ be the postchange density, where $\theta \in \Theta$. Denote $Z_n(\theta) =$

$\log[f_\theta(X_n)/f_0(X_n)]$. Then, inequality (3.9) in Theorem 3.1 holds for expectation $\mathsf{E}_\theta[\tau^*]$. Additionally, assuming that the Kullback-Leibler information number $I(\theta) = \mathsf{E}_\theta[Z_1(\theta)]$ is positive and finite, the asymptotic lower bound given by equation (3.10) holds with $I(\theta)$; that is,

$$\inf_{T \in \mathbb{C}(\gamma)} \mathsf{ESEDD}_\theta(T) \geqslant \frac{\log \gamma}{I(\theta)}(1 + o(1)) \quad \text{as } \gamma \to \infty, \tag{3.12}$$

where $\mathsf{ESEDD}_\theta(T) = \sup_{0 \leqslant \nu < \infty} \mathrm{ess\,sup}\, \mathsf{E}_{\nu,\theta}[(T-\nu)^+|\mathscr{F}_\nu]$ and $\mathsf{E}_{\nu,\theta}$ is the expectation under $\mathsf{P}_{\nu,\theta}$ when the change occurs at ν with the postchange density f_θ.

Lorden (1971) addressed the composite hypothesis for the exponential family given by equation (2.30) with $f_0 = f_{\theta=0}$; that is,

$$\frac{f_\theta(X_n)}{f_0(X_n)} = \exp\left\{\theta X_n - b(\theta)\right\}, \quad \theta \in \Theta, \quad n = 1, 2, ...,$$

where $b(\theta)$ is a convex and infinitely differentiable function on the natural parameter space Θ, $b(0) = 0$. Let $\tilde{\Theta} = \Theta - 0$.

To find asymptotically optimal procedures by applying Theorem 3.1 along with inequality (3.12), we need to determine stopping times, $\tau(\gamma) \in \mathbb{C}(\gamma)$, such that

$$\mathsf{P}_0(\tau(\gamma) < \infty) \leqslant 1/\gamma \quad \text{for } \gamma > 0 \tag{3.13}$$

and

$$\mathsf{E}_\theta[\tau(\gamma)] = \frac{\log \gamma}{I(\theta)}(1 + o(1)) \quad \text{as } \gamma \to \infty \quad \text{for all } \theta \in \tilde{\Theta}, \tag{3.14}$$

where $I(\theta) = \theta \dot{b}(\theta) - b(\theta)$.

The LLR for the sample $(X_{k+1}, ..., X_n)$ is

$$\lambda_n^k(\theta) =: \log\left[\prod_{t=k+1}^n \frac{f_\theta(X_t)}{f_0(X_t)}\right] = \theta S_n^k - (n-k)b(\theta),$$

where $S_n^k = X_{k+1} + \cdots + X_n$. Define the GLR one-sided test

$$\tau(h) = \inf\left\{n \geqslant 1 : \sup_{\theta \geqslant |\theta_1|}\left[\theta S_n^0 - nb(\theta)\right] > h(\gamma)\right\},$$

where θ_1 may be either a fixed value if the alternative hypothesis is $\theta \leqslant -\theta_1$ or $\theta \geqslant \theta_1$ or $\theta_1(\gamma) \to 0$ as $\gamma \to \infty$ if the hypothesis is $\theta \neq 0$. Lorden demonstrated that

$$\mathsf{P}_0(\tau(h) < \infty) \leqslant \exp\left\{-h(\gamma)\right\}\left[1 + \frac{h(\gamma)}{\min(I(\theta_1), I(-\theta_1))}\right], \tag{3.15}$$

so $h(\gamma)$ can be selected so that $h(\gamma) \sim \log \gamma$ as $\gamma \to \infty$. Hence, equations (3.13) and (3.14) hold. Applying $\tau(h)$ to $X_{k+1}, X_{k+2}, ...$, we obtain the stopping time $\tau_k(h)$, so that $\tau^*(h) = \min_{k \geqslant 0}(\tau_k + k)$ is the stopping time of the GLR CUSUM procedure,

$$\tau^*(h) = \inf\left\{n \geqslant 1 : \max_{0 \leqslant \nu \leqslant n} \sup_{\theta \geqslant |\theta_1|}\left[\theta S_n^\nu - (n-\nu)b(\theta)\right] > h(\gamma)\right\}.$$

Thus, the GLR CUSUM procedure is asymptotically first-order minimax.

The inequality (3.15) is usually overly pessimistic. A much better result gives the approximation $P_0(\tau(h) < \infty) \approx \sqrt{h(\gamma)} \exp\{-h(\gamma)\}C$, which follows from equation (2.44). However, the latter one does not guarantee the inequality $P_0(\tau(h) < \infty) \leqslant \gamma^{-1}$, and therefore the inequality $E_\infty[\tau^*(h)] \geqslant \gamma$.

Later, Lorden and Pollak (2005, 2008) proposed adaptive Shiryaev-Roberts and CUSUM procedures that utilize one-step delayed estimators of unknown postchange parameters θ. In these procedures, an estimate $\hat{\theta}_{n-1}(X_1, ..., X_{n-1})$ is used after observing the sample of size n, similar to the Robbins-Siegmund one-sided adaptive SPRT; see Robbins and Siegmund (1972, 1974). They compared the performance of these adaptive procedures with that of the mixture-based Shiryaev-Roberts procedure. Notably, these adaptive procedures are computationally simpler than the GLR CUSUM procedure.

We conclude with some remarks on later, related developments.

4. REMARKS

1. Fifteen years later, Moustakides (1986) advanced Lorden's asymptotic theory by demonstrating, using optimal stopping theory, that the CUSUM procedure is strictly optimal for any ARL to false alarm $\gamma \geqslant 1$ if the threshold $a = a(\gamma)$ is chosen such that $\mathsf{ARLFA}(T_a) = \gamma$.
2. Shiryaev (1996) showed that the CUSUM procedure is strictly optimal in the continuous-time scenario for detecting the change in the mean of the Wiener process according to Lorden's minimax criterion.
3. Pollak (1985) introduced a distinct minimax criterion aimed at minimizing the supremum expected detection delay $\sup_{\nu \geqslant 0} E[T - \nu | T > \nu]$. Additionally, Pollak proposed a modification of the conventional Shiryaev-Roberts (SR) procedure known as the SRP procedure, which initiates from a randomly distributed point following the quasi-stationary distribution of the SR statistic. He proved that this procedure is third-order asymptotically minimax, minimizing $\sup_{\nu \geqslant 0} E[T - \nu | T > \nu]$ to within $o(1)$ as $\gamma \to \infty$ within the class $\mathbb{C}(\gamma)$.
4. Tartakovsky, Pollak, and Polunchenko (2012) proved that the specially designed SR-r procedure that starts from a fixed point $r = r(\gamma)$ is third-order asymptotically optimal with respect to Pollak's measure $\sup_{\nu \geqslant 0} E[T - \nu | T > \nu]$ within the class $\mathbb{C}(\gamma)$ as $\gamma \to \infty$.
5. Polunchenko and Tartakovsky (2010) demonstrated that the specially designed SR-r procedure, which commences from a predetermined point $r = r(\gamma)$, is strictly optimal with respect to Pollak's measure $\sup_{\nu \geqslant 0} E[T - \nu | T > \nu]$ within the class $\mathbb{C}(\gamma)$ for a specific model.
6. Pollak and Tartakovsky (2009) proved strict optimality of the repeated SR procedure that starts from zero in the problem of detecting distant changes.
7. Moustakides, Polunchenko, and Tartakovsky (2009) conducted a thorough comparison of CUSUM and SR procedures, demonstrating that CUSUM outperforms SR in terms of the conditional expected detection delay $E_\nu[T - \nu | T > \nu]$ for relatively small values of the change point ν. However, SR proves to be more effective than CUSUM for relatively large ν.

ACKNOWLEDGEMENTS

DISCLOSURE STATEMENT

REFERENCES

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: Statistical Methodology* 12 (1):137–44. https://doi.org/10.1111/j.2517-6161.1950.tb00050.x.

Bartroff, J. 2006a. "Efficient Three-Stage *t*-Tests." In *Recent Developments in Nonparametric Inference and Probability: Festschrift for Michael Woodroofe*. Vol. 50 of *IMS Lecture Notes Monograph Series*, edited by Michael Woodroofe and Jiayang Sun, 105–11. Hayward: Institute of Mathematical Statistics.

Bartroff, J. 2006b. "Optimal Multistage Sampling in a Boundary-Crossing Problem." *Sequential Analysis* 25 (1):59–84. https://doi.org/10.1080/07474940500452247.

Bartroff, J. 2007. "Asymptotically Optimal Multistage Tests of Simple Hypotheses." *The Annals of Statistics* 35 (5):2075–105. https://doi.org/10.1214/009053607000000235.

Bartroff, J., and T. L. Lai. 2008a. "Efficient Adaptive Designs with Mid-Course Sample Size Adjustment in Clinical Trials." *Statistics in Medicine* 27 (10):1593–611. https://doi.org/10.1002/sim.3201.

Bartroff, J., and T. L. Lai. 2008b. "Generalized Likelihood Ratio Statistics and Uncertainty Adjustments in Adaptive Design of Clinical Trials." *Sequential Analysis* 27 (3):254–76. https://doi.org/10.1080/07474940802241009.

Bartroff, J., T. L. Lai, and M. Shih. 2013. *Sequential Experimentation in Clinical Trials: Design and Analysis*. New York: Springer.

Chernoff, Herman. 1959. "Sequential Design of Experiments." *The Annals of Mathematical Statistics* 30 (3):755–70. https://doi.org/10.1214/aoms/1177706205.

Devlin, Keith, and Gary Lorden. 2007. *The Numbers Behind NUMB3RS: Solving Crime With Mathematics*. New York: Penguin Books.

Dragalin, V. P., and A. A. Novikov. 1987. "Asymptotic Solution of the Kiefer–Weiss Problem for Processes with Independent Increments." *Theory of Probability & Its Applications* 32 (4):617–27. https://doi.org/10.1137/1132094.

Feller, William. 1966. *An Introduction to Probability Theory and Its Applications*. 2nd ed. Vol. 2 of *Series in Probability and Mathematical Statistics*. New York, NY: John Wiley & Sons, Inc.

Fellouris, Georgios, and Alexander G. Tartakovsky. 2017. "Multichannel Sequential Detection—Part I: Non-i.i.d. data." *IEEE Transactions on Information Theory* 63 (7):4551–71. https://doi.org/10.1109/TIT.2017.2689785.

Huffman, M. D. 1983. "An Efficient Approximate Solution to the Kiefer–Weiss Problem." *Annals of Statistics* 11 (1):306–16.

Kalashnikov, Vladimir V. 2013. *Mathematical Methods in Queuing Theory*. Vol. 271. Dordrecht: Springer Science & Business Media.

Kiefer, Jack, and J. Sacks. 1963. "Asymptotically Optimal Sequential Inference and Design." *The Annals of Mathematical Statistics* 34 (3):705–50. https://doi.org/10.1214/aoms/1177704000.

Kiefer, Jack, and Lionel Weiss. 1957. "Properties of Generalized Sequential Probability Ratio Tests." *The Annals of Mathematical Statistics* 28 (1):57–74. https://doi.org/10.1214/aoms/1177707037.

Lai, Tze Leung. 1981. "Asymptotic Optimality of Invariant Sequential Probability Ratio Tests." *Annals of Statistics* 9 (2):318–33.

Lai, Tze Leung. 1998. "Information Bounds and Quick Detection of Parameter Changes in Stochastic Systems." *IEEE Transactions on Information Theory* 44 (7):2917–29.

Lai, Tze Leung, and Mei-Chiung Shih. 2004. "Power, Sample Size and Adaptation Considerations in the Design of Group Sequential Clinical Trials." *Biometrika* 91 (3):507–28. https://doi.org/10.1093/biomet/91.3.507.

Lorden, Gary. 1967. "Integrated Risk of Asymptotically Bayes Sequential Tests." *The Annals of Mathematical Statistics* 38 (5):1399–422. https://doi.org/10.1214/aoms/1177698696.

Lorden, Gary. 1970. "On Excess over the Boundary." *The Annals of Mathematical Statistics* 41 (2):520–7. https://doi.org/10.1214/aoms/1177697092.

Lorden, Gary. 1971. "Procedures for Reacting to a Change in Distribution." *The Annals of Mathematical Statistics* 42 (6):1897–908. https://doi.org/10.1214/aoms/1177693055.

Lorden, Gary. 1972. "Likelihood Ratio Tests for Sequential k-Decision Problems." *The Annals of Mathematical Statistics* 43 (5):1412–27. https://doi.org/10.1214/aoms/1177692374.

Lorden, Gary. 1973. "Open-Ended Tests for Koopman-Darmois Families." *Annals of Statistics* 1 (4):633–43.

Lorden, Gary. 1976. "2-SPRT's and the Modified Kiefer-Weiss Problem of Minimizing an Expected Sample Size." *Annals of Statistics* 4 (2):281–91.

Lorden, Gary. 1977a. "Nearly optimal sequential tests for exponential families." *Unpublished Manuscript*. Available from http://jaybartroff.com/research/gary.pdf

Lorden, Gary. 1977b. "Nearly-Optimal Sequential Tests for Finitely Many Parameter Values." *Annals of Statistics* 5 (1):1–21.

Lorden, Gary. 1980. "Structure of Sequential Tests Minimizing an Expected Sample Size." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 51 (3):291–302. https://doi.org/10.1007/BF00587355.

Lorden, Gary. 1983. "Asymptotic Efficiency of Three-Stage Hypothesis Tests." *The Annals of Statistics* 11 (1):129–40. https://doi.org/10.1214/aos/1176346064.

Lorden, Gary, and Moshe Pollak. 2005. "Nonanticipating Estimation Applied to Sequential Analysis and Changepoint Detection." *Annals of Statistics* 33 (3):1422–54.

Lorden, Gary, and Moshe Pollak. 2008. "Sequential Change-Point Detection Procedures That Are Nearly Optimal and Computationally Simple." *Sequential Analysis* 27 (4):476–512. https://doi.org/10.1080/07474940802446244.

Moustakides, George V. 1986. "Optimal Stopping Times for Detecting Changes in Distributions." *Annals of Statistics* 14 (4):1379–87.

Moustakides, George V., Aleksey S. Polunchenko, and Alexander G. Tartakovsky. 2009. "Numerical Comparison of CUSUM and Shiryaev–Roberts Procedures for Detecting Changes in Distributions." *Communications in Statistics - Theory and Methods* 38 (16-17):3225–39. https://doi.org/10.1080/03610920902947774.

Nathan, A., J. Albert, J. Bartroff, R. Blandford, D. Brooks, J. Derenski, L. Goldstein, P. Hosoi, G. Lorden, and L. Smith. 2018. Report of the committee studying home run rates in major league baseball. Technical report, Commissioner of Major League Baseball.

Novak, Serguei Y. 2011. *Extreme Value Methods with Applications to Finance.* Boca Raton, FL: CRC Press.

Page, E. S. 1954. "Continuous Inspection Schemes." *Biometrika* 41 (1-2):100–15. https://doi.org/10.1093/biomet/41.1-2.100.

Pollak, Moshe. 1985. "Optimal Detection of a Change in Distribution." *Annals of Statistics* 13 (1):206–27.

Pollak, M., and A. G. Tartakovsky. 2009. "Optimality Properties of the Shiryaev–Roberts Procedure." *Statistica Sinica* 19 (4):1729–39.

Polunchenko, A. S., and A. G. Tartakovsky. 2010. "On Optimality of the Shiryaev–Roberts Procedure for Detecting a Change in Distribution." *Annals of Statistics* 38 (6):3445–57.

Rausand, Marvin, and Arnljot Hoyland. 2003. *System Reliability Theory: Models, Statistical Methods, and Applications*. Vol. 396. Hoboken, NJ: John Wiley & Sons.

Robbins, Herbert, and David Siegmund. 1972. "A Class of Stopping Rules for Testing Parameter Hypotheses." In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, June 21–July 18, 1970*. Vol. 4 of *Biology and Health*, edited by L. M. Le Cam, J. Neyman, and E. L. Scott, 37–41. Berkeley, CA: University of California Press.

Robbins, Herbert, and David Siegmund. 1974. "The Expected Sample Size of Some Tests of Power One." *Annals of Statistics* 2 (3):415–36.

Schwarz, G. 1962. "Asymptotic Shapes of Bayes Sequential Testing Regions." *The Annals of Mathematical Statistics* 33 (1):224–36. https://doi.org/10.1214/aoms/1177704726.

Shewhart, WalterAndrew. 1931. *Economic Control of Quality of Manufactured Products*. New York: D. Van Nostrand Co.

Shiryaev, A. N. 1996. "Minimax Optimality of the Method of Cumulative Sums (CUSUM) in the Case of Continuous Time." *Russian Mathematical Surveys* 51 (4):750–1. https://doi.org/10.1070/RM1996v051n04ABEH002986.

Siegmund, David. 1985. "*Sequential Analysis: Tests and Confidence Intervals*." Series in Statistics. New York: Springer-Verlag.

Tartakovsky, A. G. 1998. "Asymptotic Optimality of Certain Multihypothesis Sequential Tests: Non-i.i.d. Case." *Statistical Inference for Stochastic Processes* 1 (3):265–95. https://doi.org/10.1023/A:1009952514505.

Tartakovsky, A. G. 2020. *Sequential Change Detection and Hypothesis Testing: General Non-i.i.d. Stochastic Models and Asymptotically Optimal Rules. Monographs on Statistics and Applied Probability 165*. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group.

Tartakovsky, A. G. 2024. "Nearly Optimum Properties of Certain Multi-Decision Sequential Rules for General Non-i.i.d. Stochastic Models." *Annals of Mathematical Sciences and Applications* (submitted). arXiv2405.00928 available at http://arxiv.org/abs/2405.00928.

Tartakovsky, A. G., I. V. Nikiforov, and M. Basseville. 2015. *Sequential Analysis: Hypothesis Testing and Changepoint Detection. Monographs on Statistics and Applied Probability 136*. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group.

Tartakovsky, Alexander G., Moshe Pollak, and Aleksey S. Polunchenko. 2012. "Third-Order Asymptotic Optimality of the Generalized Shiryaev–Roberts Changepoint Detection Procedures." *Theory of Probability & Its Applications* 56 (3):457–84. https://doi.org/10.1137/S0040585X97985534.

Wald, Abraham. 1945. "Sequential Tests of Statistical Hypotheses." *The Annals of Mathematical Statistics* 16 (2):117–86. https://doi.org/10.1214/aoms/1177731118.

Wald, Abraham. 1946. "Differentiation Under the Expectation Sign in the Fundamental Identity of Sequential Analysis." *The Annals of Mathematical Statistics* 17 (4):493–7. https://doi.org/10.1214/aoms/1177730889.

Wald, Abraham. 1947. *Sequential Analysis*. New York: John Wiley & Sons, Inc.

Wald, Abraham, and J. Wolfowitz. 1948. "Optimum Character of the Sequential Probability Ratio Test." *The Annals of Mathematical Statistics* 19 (3):326–39. https://doi.org/10.1214/aoms/1177730197.

Weiss, Lionel. 1962. "On Sequential Tests Which Minimize the Maximum Expected Sample Size." *Journal of the American Statistical Association* 57 (299):551–66. https://doi.org/10.1080/01621459.1962.10500543.

Whitehead, John. 1997. *The Design and Analysis of Sequential Clinical Trials*. Chichester, UK: John Wiley & Sons.

Wong, Seok Pin. 1968. "Asymptotically Optimum Properties of Certain Sequential Tests." *The Annals of Mathematical Statistics* 39 (4):1244–63. https://doi.org/10.1214/aoms/1177698250.

Xing, Yiming, and Georgios Fellouris. 2023. "Signal Recovery with Multistage Tests and without Sparsity Constraints." *IEEE Transactions on Information Theory* 69 (11):7220–45. https://doi.org/10.1109/TIT.2023.3299874.