Article

Quick and Complete Convergence in the Law of Large Numbers with Applications to Statistics

Alexander Tartakovsky *

Posted Date: 12 May 2023

doi: 10.20944/preprints202305.0904.v1

Keywords: Complete convergence; r-quick convergence; sequential analysis; hypothesis testing; changepoint detection

Article

Quick and Complete Convergence in the Law of Large Numbers with Applications to Statistics

Alexander G Tartakovsky †

AGT StatConsult; alexg.tartakovsky@gmail.com
* Correspondence: alexg.tartakovsky@gmail.com; Tel.: +1-310-292-7847
† Current address: 71 Cypress Way, Rolling Hills Estates, California 90274, USA

Abstract: In the first part of this article, we discuss and generalize the complete convergence introduced by Hsu and Robbins (1947) to r-complete convergence introduced by Tartakovsky (1998). We also establish its relation to the r-quick convergence first introduced by Strassen (1967) and extensively studied by Lai (1976). Our work is motivated by various statistical problems, mostly in sequential analysis. As we show in the second part, generalizing and studying these convergence modes is important not only in probability theory but also to solve challenging statistical problems in hypothesis testing and changepoint detection for general stochastic non-i.i.d. models.

Keywords: Complete convergence; r-quick convergence; sequential analysis; hypothesis testing; changepoint detection

1. Introduction

In [1], Hsu and Robbins introduced the notion of complete convergence which is stronger than almost sure (a.s.) convergence. Hsu and Robbins used this notion to discuss certain aspects of the Law of Large Numbers (LLN). In particular, let $X_1, X_2, \dots$ be independent and identically distributed (i.i.d.) random variables with the common mean $\mu = \mathsf{E}[X_1]$. Hsu and Robbins proved that, while in the Kolmogorov Strong Law of Large Numbers (SLLN), only the first moment condition is needed for the sample mean $n^{-1}\sum_{t=1}^n X_t$ to converge to μ as $n \to \infty$, the complete version of the SLLN requires the second-moment condition $\mathsf{E}|X_1|^2 < \infty$ (finiteness of variance). Later, Baum and Katz [2], working on the rate of convergence in the LLN, established that the second-moment condition is not only necessary but also sufficient for complete convergence. Strassen [3] introduced another mode of convergence, the r-quick convergence. When $r = 1$, these two modes of convergence are closely related. In the case of i.i.d. random variables and the sample mean $n^{-1}\sum_{t=1}^n X_t$, they are identical. This fact and certain statistical applications motivated Tartakovsky [4] (see also Tartakovsky [5] and Tartakovsky et al. [6]) to introduce a natural generalization of complete convergence – the r-complete convergence, which turns out to be identical to the r-quick convergence in the i.i.d. case.

Section 2 discusses pure probabilistic issues related to r-complete convergence and r-quick convergence. Section 3 explores statistical applications in sequential hypothesis testing and changepoint detection. Section 4 outlines sufficient conditions for r-complete convergence for Markov and hidden Markov models, which is needed to establish optimality properties of sequential hypothesis tests and changepoint detection procedures. Section 5 concludes.

2. Modes of Convergence and the Law of Large Numbers

We begin by listing some standard definitions in probability theory. Let $(\Omega, \mathscr{F})$ be a measurable space, i.e., Ω is a set of elementary events ω and $\mathscr{F}$ is a sigma-algebra (a system of subsets of Ω satisfying standard conditions). A probability space is a triple $(\Omega, \mathscr{F}, \mathsf{P})$, where P is a probability measure (completely additive measure normalized to 1) defined on the sets from the sigma-algebra $\mathscr{F}$. More specifically, by Kolmogorov's axioms, probability P satisfies: $\mathsf{P}(\mathcal{A}) \ge 0$ for any $\mathcal{A} \in \mathscr{F}$; $\mathsf{P}(\Omega) = 1$; and $\mathsf{P}(\cup_{i=1}^\infty \mathcal{A}_i) = \sum_{i=1}^\infty \mathsf{P}(\mathcal{A}_i)$ for $\mathcal{A}_i \in \mathscr{F}$, $\mathcal{A}_i \cap \mathcal{A}_j = \varnothing$, $i \neq j$, where $\varnothing$ is an empty set.

A function $X = X(\omega)$ defined on $(\Omega, \mathscr{F})$ with values in $\mathscr{X}$ is called random variable if it is $\mathscr{F}$-measurable, i.e., $\{\omega : X(\omega) \in B\}$ belongs to the sigma-algebra $\mathscr{F}$. The function $F(x) = \mathsf{P}(\omega : X(\omega) \le x)$ is the distribution function of X. It is also referred to as cumulative distribution function (cdf). The real-valued random variables $X_1, X_2, \dots$ are independent if the events $\{X_1 \le x_1\}, \{X_2 \le x_2\}, \dots$ are independent for every sequence $x_1, x_2, \dots$ of real numbers. In what follows, we shall deal with real-valued random variables unless specified otherwise.

2.1. Standard Modes of Convergence

Let X be a random variable and let $\{X_n\}_{n\in\mathbb{Z}_+}$ ($\mathbb{Z}_+ = \{0, 1, 2, \dots\}$) be a sequence of random variables, both defined on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$. We now give several standard definitions and results related to the Law of Large Numbers.

Convergence in Distribution (Weak Convergence). Let $F_n(x) = \mathsf{P}(\omega : X_n \le x)$ be the cdf of X_n and let $F(x) = \mathsf{P}(\omega : X \le x)$ be the cdf of X. We say that the sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to X in distribution (or in law or weakly) as $n \to \infty$ and write $X_n \xrightarrow[n\to\infty]{\text{law}} X$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all continuity points of $F(x)$.

Convergence in Probability. We say that the sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to X in probability as $n \to \infty$ and write $X_n \xrightarrow[n\to\infty]{\mathsf{P}} X$ if

$$\lim_{n\to\infty} \mathsf{P}(|X_n - X| > \varepsilon) = 0 \quad \text{for every } \varepsilon > 0.$$

Almost Sure Convergence. We say that the sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to X almost surely (a.s.) or with probability 1 (w.p. 1) as $n \to \infty$ under probability measure P and write $X_n \xrightarrow[n\to\infty]{\mathsf{P}-\text{a.s.}} X$ if

$$\mathsf{P}\left(\omega : \lim_{n\to\infty} X_n = X\right) = 1. \tag{1}$$

It is easily seen that (1) is equivalent to the condition

$$\lim_{n\to\infty} \mathsf{P}\left(\omega : \sum_{t=n}^{\infty} |X_t - X| > \varepsilon\right) = 0 \quad \text{for every } \varepsilon > 0,$$

and that the a.s. convergence implies convergence in probability, and the convergence in probability implies convergence in distribution, while the converse statements are not generally true.

The following double implications that establish necessary and sufficient conditions (i.e., equivalences) for the a.s. convergence are useful:

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \quad \Longleftrightarrow \quad \mathsf{P}\left(\sup_{t\ge n} |X_t - X| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \text{for all } \varepsilon > 0. \tag{2}$$

The following result is often useful.

Lemma 1. *Let $f(t)$ be a nonnegative increasing function, $\lim_{t\to\infty} f(t) = \infty$. If*

$$\frac{X_n}{f(n)} \xrightarrow[n\to\infty]{\mathsf{P}-a.s.} 0,$$

then

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n} X_t > \varepsilon\right) = 0 \quad \textit{for every } \varepsilon > 0. \tag{3}$$

Proof. For any $\varepsilon > 0$, $n_0 > 0$ and $n > n_0$, we have

$$\begin{aligned}\mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n} X_t > \varepsilon\right) &\le \mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n_0} X_t > \varepsilon\right) + \mathsf{P}\left(\frac{1}{f(n)} \max_{n_0< t\le n} X_t > \varepsilon\right) \\ &\le \mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n_0} X_t > \varepsilon\right) + \mathsf{P}\left(\max_{t>n_0} \frac{X_t}{f(t)} > \varepsilon\right).\end{aligned}$$

Letting $n \to \infty$ and taking into account that

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n_0} X_t > \varepsilon\right) = 0,$$

we obtain

$$\limsup_{n\to\infty} \mathsf{P}\left(\frac{1}{f(n)} \max_{0\le t\le n} X_t > \varepsilon\right) \le \mathsf{P}\left(\sup_{t>n_0} \frac{X_t}{f(t)} > \varepsilon\right).$$

Since n_0 can be arbitrarily large, we can let $n_0 \to \infty$ and since, by assumption $X_n/f(n) \xrightarrow[n\to\infty]{\text{a.s.}} 0$, it follows from (2) that the upper bound approaches 0 as $n_0 \to \infty$. This completes the proof. □

Remark 1. *The proof of Lemma 1 shows that the assertion (3) also holds under the one-sided condition*

$$\mathsf{P}\left(\sup_{t>n} \frac{X_t}{f(t)} > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \textit{for all } \varepsilon > 0. \tag{4}$$

Random Walk. Let $X_0, X_1, X_2, \dots$ be i.i.d. random variables with mean $\mathsf{E}[X_n] = \mu$ for $n \ge 1$ and the initial condition $X_0 = x$. Then $S_n = \sum_{t=0}^{n} X_t$ is called a random walk with mean $x + \mu\, n$.

In what follows, in the case where $X_1, X_2, \dots$ are i.i.d. random variables and $S_n = \sum_{t=0}^{n} X_t$, we prefer to formulate the results in terms of the random walk $\{S_n\}_{n\in\mathbb{Z}_+}$ (typically $S_0 = 0$ while not necessarily).

We now recall the two Strong Law of Large Numbers (SLLN). Write $S_n = X_0 + X_1 + \cdots + X_n$ for the partial sum ($X_0 = S_0 = 0$), so that $\{S_n\}_{n\in\mathbb{Z}_+}$ is a random walk with zero initial condition as long as $X_1, X_2, \dots$ are i.i.d. with mean μ.

Kolmogorov's SLLN. Let $\{S_n\}_{n\in\mathbb{Z}_+}$ be a random walk under probability measure P. If $\mathsf{E}[S_1]$ exists, then the sample mean S_n/n converges to the mean value $\mathsf{E}[S_1]$ w.p. 1, i.e.,

$$n^{-1} S_n \xrightarrow[n\to\infty]{\mathsf{P}-\text{a.s.}} \mathsf{E}[S_1]. \tag{5}$$

Conversely, if $n^{-1} S_n \xrightarrow[n\to\infty]{\mathsf{P}-\text{a.s.}} \mu$, where $|\mu| < \infty$, then $\mathsf{E}[S_1] = \mu$.

Marcinkiewicz-Zygmund's SLLN. Let $\{S_n\}_{n\in\mathbb{Z}_+}$ be a zero-mean random walk under probability measure P. The following two statements are equivalent:

(i) $\mathsf{E}\,|S_1|^p < \infty$ for $0 < p < 2$;

(ii) $n^{-1/p} S_n \xrightarrow[n\to\infty]{\mathsf{P}-\text{a.s.}} 0$.

2.2. Complete and r-Complete Convergence

We begin with discussing the issue of rates of convergence in the LLN.

doi:10.20944/preprints202305.0904.v1

Rates of Convergence. Let $\{X_n\}_{n\in\mathbb{Z}_+}$ be a sequence of random variables and assume that X_n converges to 0 w.p. 1 as $n \to \infty$. The question is what the rate of convergence is? In other words, how fast does the tail probability $\mathsf{P}(|X_n| > \varepsilon)$ decay to zero? This question can be answered by analyzing the behavior of the sums

$$\Sigma(r,\varepsilon) := \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}(|X_n| > \varepsilon) \quad \text{for some } r > 0 \text{ and all } \varepsilon > 0.$$

More specifically, if $\Sigma(r,\varepsilon)$ is finite for every $\varepsilon > 0$, then the tail probability $\mathsf{P}(|X_n| > \varepsilon)$ decays with the rate faster than $1/n^r$, so that $n^r\mathsf{P}(|X_n| > \varepsilon) \to 0$ for all $\varepsilon > 0$ as $n \to \infty$.

To answer these questions we now consider modes of convergence that strengthen the almost sure convergence, and therefore, help to determine the rate of convergence in the SLLN. Historically this issue was first addressed in 1947 by Hsu and Robbins [1] who introduced the new mode of convergence that they called *Complete Convergence*.

Complete Convergence. The sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to 0 *completely* if

$$\lim_{n\to\infty} \sum_{i=n}^{\infty} \mathsf{P}(|X_t| > \varepsilon) = 0 \quad \text{for every } \varepsilon > 0. \tag{6}$$

Clearly, (6) is equivalent to

$$\Sigma(1,\varepsilon) = \sum_{n=1}^{\infty} \mathsf{P}(|X_n| > \varepsilon) < \infty \quad \text{for every } \varepsilon > 0.$$

Also, (6) implies a.s. convergence $X_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$, but converse is not generally true unless the variables $X_1, X_2, \dots$ are not independent.

Let $\{S_n\}_{n\in\mathbb{Z}_+}$ be a random walk with mean $\mathsf{E}[S_n] = \mu\, n$. Kolmogorov's SLLN (5) implies that the sample mean S_n/n converges to μ w.p. 1. Hsu and Robbins [1] proved that under the same assumptions (i.e., under the only first-moment condition $\mathsf{E}|S_1| < \infty$) the sequence $\{n^{-1}S_n\}_{n\ge 1}$ need not converge to μ completely, but it will do so under the further second-moment condition $\mathsf{E}|S_1|^2 < \infty$. So the finiteness of variance is a sufficient condition for complete convergence in the SLLN. They conjectured that the second-moment condition is not only sufficient but also necessary for complete convergence. Thus, it follows from these results that if the variance is finite, then the rate of convergence in Kolmogorov's SLLN is $\lim_{n\to\infty} n\,\mathsf{P}(|S_n/n - \mu| > \varepsilon) = 0$ for all $\varepsilon > 0$.

A further step towards this issue was done in 1965 by Baum and Katz [2]. In particular, the following result follows from Theorem 3 in [2] for the random walk $\{S_n\}_{n\in\mathbb{Z}_+}$ with mean $\mathsf{E}[S_1] = \mu$.

Theorem 1. *Let $r > 0$ and $\alpha > 1/2$. If $\{S_n\}_{n\in\mathbb{Z}_+}$ is a random walk with mean $\mathsf{E}[S_1] = \mu$, then the following statements are equivalent:*

$$\begin{aligned} \mathsf{E}[|S_1|^{(r+1)/\alpha}] < \infty &\iff \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left\{\frac{1}{n^\alpha}|S_n - \mu n| > \varepsilon\right\} < \infty \ \textit{for all } \varepsilon > 0 \\ &\iff \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left\{\sup_{k\ge n}\frac{1}{k^\alpha}|S_k - \mu k| > \varepsilon\right\} < \infty \ \textit{for all } \varepsilon > 0. \end{aligned} \tag{7}$$

Setting $r = 1$ and $\alpha = 1$ in (7), we obtain the following equivalence

$$\mathsf{E}[|S_1|^2] < \infty \iff \sum_{n=1}^{\infty} \mathsf{P}\{|S_n/n - \mu| > \varepsilon\} \quad \text{for all } \varepsilon > 0,$$

which shows that the conjecture of Hsu and Robbins is correct – the second-moment condition $\mathsf{E}|S_1|^2 < \infty$ is both necessary and sufficient for complete convergence

$$n^{-1}S_n \xrightarrow[n\to\infty]{\mathsf{P}-\text{completely}} \mu.$$

Furthermore, if for some $r > 0$ the $(r+1)$-th moment is finite, $\mathsf{E}|S_1|^{r+1} < \infty$, then the rate of convergence in the SLLN is $\lim_{n\to\infty} n^r\,\mathsf{P}(|S_n/n-\mu|>\varepsilon)=0$ for all $\varepsilon>0$.

Previous results suggest that it is reasonable to generalize the notion of complete convergence into the following mode of convergence that we will refer to as *r-Complete Convergence*, which is also related to the so-called r-Quick Convergence that we will discuss later on (see Subsection 2.3).

Definition 1 (***r*-Complete Convergence**)**.** *Let $r > 0$. We say that the sequence of random variables $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to X r-completely as $n\to\infty$ under probability measure* P *and write $X_n \xrightarrow[n\to\infty]{\mathsf{P}\text{-}r\text{-completely}} X$ if*

$$\Sigma(r,\varepsilon) := \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}(|X_n-X|>\varepsilon)<\infty \quad \textit{for every } \varepsilon>0. \tag{8}$$

Note that the a.s. convergence of $\{X_n\}$ to X can be equivalently written as

$$\lim_{n\to\infty}\mathsf{P}\left(\sum_{i=n}^{\infty}|X_t-X|>\varepsilon\right)=0 \quad \text{for every } \varepsilon>0,$$

so that the r-complete convergence with $r\ge 1$ implies the a.s. convergence, but the converse is not true in general.

Suppose that X_n converges a.s. to X. If $\Sigma(r,\varepsilon)$ is finite for every $\varepsilon>0$, then

$$\lim_{n\to\infty}\sum_{t=n}^{\infty} t^{r-1}\mathsf{P}(|X_t-X|>\varepsilon)=0 \quad \text{for every } \varepsilon>0$$

and probability $\mathsf{P}(|X_n-X|>\varepsilon)$ goes to 0 as $n\to\infty$ with the rate faster than $1/n^r$. Hence, as already mentioned above, the r-complete convergence allows one to determine the rate of convergence of X_n to X, i.e., to answer the question on how fast the tail probability $\mathsf{P}(|X_n-X|>\varepsilon)$ decays to zero.

The following result provides a very useful implication of complete convergence.

Theorem 2. *Let $\{X_n\}_{n\in\mathbb{Z}_+}$ and $\{Y_n\}_{n\in\mathbb{Z}_+}$ be two arbitrary, possibly dependent sequences of random variables. Assume that there are positive and finite numbers μ_1 and μ_2 such that*

$$\sum_{n=1}^{\infty}\mathsf{P}\left(\left|\frac{1}{n}X_n-\mu_1\right|>\varepsilon\right)<\infty \quad \textit{for every } \varepsilon>0 \tag{9}$$

and

$$\sum_{n=1}^{\infty}\mathsf{P}\left(\left|\frac{1}{n}Y_n-\mu_2\right|>\varepsilon\right)<\infty \quad \textit{for every } \varepsilon>0, \tag{10}$$

i.e., $n^{-1}X_n \xrightarrow[n\to\infty]{\mathsf{P}-completely} \mu_1$ and $n^{-1}Y_n \xrightarrow[n\to\infty]{\mathsf{P}-completely} \mu_2$. If $\mu_1\ge\mu_2$, then for any random time T

$$\mathsf{P}\left(X_T<b,\, Y_{T+1}\ge b(1+\delta)\right)\longrightarrow 0 \quad \textit{as } b\to\infty \quad \textit{for any } \delta>0. \tag{11}$$

Proof. Fix $\delta > 0$, $c \in (0, \delta)$ and let $N_b = \lceil 1 + c)b/\mu_2 \rceil$ be the smallest integer that is larger than or equal to $(1+c)b/\mu_2$. Observe that

$$\begin{aligned}\mathsf{P}\left(X_T < b,\, Y_{T+1} \geq b(1+\delta)\right) &\leq \mathsf{P}\left(X_T \leq b,\, T \geq N_b\right) + \mathsf{P}\left(Y_{T+1} \geq (1+\delta)b,\, T < N_b\right) \\ &\leq \mathsf{P}\left(X_T \leq b,\, T \geq N_b\right) + \mathsf{P}\left(\max_{1 \leq n \leq N_b} Y_n \geq (1+\delta)b\right).\end{aligned}$$

Thus, to prove (11) it suffices to show that the two terms on the right-hand side go to 0 as $b \to \infty$.

For the first term, we notice that for any $n \geq N_b$,

$$\frac{b}{n} \leq \frac{b}{N_b} \leq \frac{\mu_2}{1+c} \leq \frac{\mu_1}{1+c} < \mu_1,$$

so that

$$\begin{aligned}\mathsf{P}\left(X_T \leq b,\, T \geq N_b\right) &= \sum_{n=N_b}^{\infty} \mathsf{P}\left(X_n \leq b,\, T = n\right) \leq \sum_{n=N_b}^{\infty} \mathsf{P}\left(\frac{X_n}{n} \leq \frac{b}{n}\right) \\ &\leq \sum_{n=N_b}^{\infty} \mathsf{P}\left(\frac{X_n}{n} \leq \frac{\mu_1}{1+c}\right) = \sum_{n=N_b}^{\infty} \mathsf{P}\left(\frac{X_n}{n} - \mu_1 \leq -\frac{c}{1+c}\mu_1\right).\end{aligned}$$

Since $N_b \to \infty$ as $b \to \infty$ the upper bound goes to 0 as $b \to \infty$ due to condition (9).

Next, since $c \in (0, \delta)$ there exists $\varepsilon' > 0$ such that

$$\frac{(1+\delta)b}{N_b} = \frac{(1+\delta)b}{\lceil b(1+c)/\mu_2 \rceil} \geq (1+\varepsilon')\mu_2.$$

As a result,

$$\mathsf{P}\left(\max_{1 \leq n \leq N_b} Y_n \geq (1+\delta)b\right) \leq \mathsf{P}\left(\frac{1}{N_b} \max_{1 \leq n \leq N_b} Y_n \geq (1+\varepsilon')\mu_2\right),$$

where the upper bound goes to 0 as $b \to \infty$ by condition (10) (see Lemma 1). □

Remark 2. *The proof suggests that the assertion* (11) *of Theorem* 2 *holds under the following one-sided conditions*

$$\mathsf{P}\left(n^{-1} \max_{1 \leq s \leq n} Y_s - \mu_2 > \varepsilon\right) \xrightarrow[n \to \infty]{} 0, \quad \sum_{n=1}^{\infty} \mathsf{P}\left(n^{-1} X_n - \mu_1 < -\varepsilon\right) < \infty.$$

Complete convergence conditions (9) *and* (10) *guarantee both these conditions.*

Remark 3. *Theorem* 2 *can be applied to the overshoot problem. Indeed, if $X_n = Y_n = Z_n$ and the random time T is the first time n when Z_n exceeds the level b, $T = \inf\{n \geq 1 : Z_n > b\}$, then Theorem* 2 *shows that the relative excess of boundary crossing (overshoot) $(Z_T - b)/b$ converges to 0 in probability as $b \to \infty$ when Z_n/n converges completely as $n \to \infty$ to a positive number μ.*

2.3. r-Quick Convergence

In 1967, Strassen [3] introduced the notion of r-quick limit points of a sequence of random variables. The r-quick convergence has been further addressed by Lai [7,8], Chow and Lai [9], Fuh and Zhang [10], and Tartakovsky [4,5] (see certain details in Subsection 2.4).

We define r-quick convergence in a way suitable for this paper. Let $\{X_n\}_{n \in \mathbb{Z}_+}$ be a sequence of real-valued random variables and let X be a random variable defined on the same probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Definition 2 (*r*-***Quick Convergence***)**.** *Let $r > 0$ and for $\varepsilon > 0$ let*

$$L_\varepsilon = \sup\{n \ge 1 : |X_n - X| > \varepsilon\} \quad (\sup\{\varnothing\} = 0)$$

be the last entry time of X_n in the region $(X + \varepsilon, \infty) \cup (-\infty, X - \varepsilon)$. We say that the sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ converges to X r-quickly as $n \to \infty$ under probability measure P and write $X_n \xrightarrow[n\to\infty]{\mathsf{P}-r\text{-quickly}} X$ if, and only if,

$$\mathsf{E}[L_\varepsilon^r] < \infty \quad \textit{for every } \varepsilon > 0, \tag{12}$$

where E is the operator of expectation under probability P.

This definition can be of course generalized to random variables X, $\{X_n\}_{n\in\mathbb{Z}_+}$ taking values in a metric space $(\mathscr{X}, d)$ with distance d: $X_n \xrightarrow[n\to\infty]{r\text{-quickly}} X$ if

$$\mathsf{E}\left[(\sup\{n \ge 1 : d(X, X_n) > \varepsilon\})^r\right] < \infty \quad \text{for every } \varepsilon > 0.$$

Note that the a.s. convergence $X_n \to \mu$ ($|\mu| < \infty$) as $n \to \infty$ to a constant μ can be expressed as $\mathsf{P}(L_\varepsilon(\mu) < \infty) = 1$, where $L_\varepsilon(\mu) = \sup\{n \ge 1 : |X_n - \mu| > \varepsilon\}$. Therefore, the r-quick convergence implies the convergence w.p. 1 but not conversely.

Note also that in general r-quick convergence is stronger than r-complete convergence. Specifically, the following lemma shows that

$$\max_{1\le i\le n} X_t \xrightarrow[n\to\infty]{r-\text{completely}} \mu \quad \Longrightarrow \quad X_n \xrightarrow[n\to\infty]{r-\text{quickly}} \mu \quad \Longrightarrow \quad X_n \xrightarrow[n\to\infty]{r-\text{completely}} \mu. \tag{13}$$

Lemma 2. *Let $\{X_n\}_{n\in\mathbb{Z}_+}$ be a sequence of random variables. Let $f(t)$ be a nonnegative increasing function, $f(0) = 0$, $\lim_{t\to\infty} f(t) = +\infty$, and let for $\varepsilon > 0$*

$$L_\varepsilon(f) = \sup\{n \ge 1 : |X_n| > \varepsilon f(n)\} \quad (\sup\{\varnothing\} = 0)$$

be the last time X_n leaves the interval $[-\varepsilon f(n), +\varepsilon f(n)]$.

(i) *For any $r > 0$ and any $\varepsilon > 0$ the following inequalities hold:*

$$r\sum_{n=1}^\infty n^{r-1}\mathsf{P}\{|X_n| \ge \varepsilon f(n)\} \le \mathsf{E}[L_\varepsilon(f)^r] \le r\sum_{n=1}^\infty n^{r-1}\mathsf{P}\left\{\sup_{t\ge n}\frac{|X_t|}{f(t)} \ge \varepsilon\right\}. \tag{14}$$

Therefore,

$$\sum_{n=1}^\infty n^{r-1}\mathsf{P}\left\{\sup_{t\ge n}\frac{|X_t|}{f(t)} \ge \varepsilon\right\} < \infty \quad \textit{for all } \varepsilon > 0 \quad \Longrightarrow \quad X_n \xrightarrow[n\to\infty]{r\text{-quickly}} 0.$$

(ii) *If $f(t)$ is a power function, $f(t) = t^\gamma$, $\gamma > 0$, then finiteness of*

$$\sum_{n=1}^\infty n^{r-1}\mathsf{P}\left\{\max_{1\le t\le n} X_t \ge \varepsilon n^\gamma\right\}$$

for some $r > 0$ and every $\varepsilon > 0$ implies r-quick convergence of X_n to 0:

$$\left\{\sum_{n=1}^\infty n^{r-1}\mathsf{P}\left(\max_{1\le t\le n} X_t \ge \varepsilon n^\gamma\right) < \infty\,\forall\,\varepsilon > 0\right\} \Longrightarrow \{\mathsf{E}[L_\varepsilon(\gamma)^r] < \infty\,\forall\,\varepsilon > 0\}, \tag{15}$$

where $L_\varepsilon(\gamma) = \sup\{n \ge 1 : |X_n| > \varepsilon\, n^\gamma\}$.

Proof. (i) Obviously,

$$\mathsf{P}\{|X_n| \geq \varepsilon f(n)\} \leq \mathsf{P}\{L_\varepsilon(f) \geq n\} \leq \mathsf{P}\left\{\sup_{t\geq n} \tfrac{1}{f(t)}\,|X_t| \geq \varepsilon\right\}$$

from which the inequalities (14) follow immediately.

(ii) Write $M_u = \max_{1\leq n\leq\lceil u\rceil} |X_n|$, where $\lceil u\rceil$ is an integer part of u. We have the following chain of inequalities and equalities:

$$\begin{aligned}
&\mathsf{E}\left[L_{2\varepsilon}(\gamma)^r\right] \leq r\int_0^\infty t^{r-1}\mathsf{P}\left\{\sup_{u\geq t} u^{-\gamma}\,|X_u| \geq 2\varepsilon\right\}\,\mathrm{d}t\\
&\leq r\int_0^\infty t^{r-1}\mathsf{P}\left\{\sup_{u\geq t}\left[|X_u| - \varepsilon u^\gamma\right] \geq \varepsilon t^\gamma\right\}\,\mathrm{d}t\\
&\leq r\int_0^\infty t^{r-1}\mathsf{P}\left\{\sup_{u>0}\left[|X_u| - \varepsilon u^\gamma\right] \geq \varepsilon t^\gamma\right\}\,\mathrm{d}t\\
&\leq r\sum_{n=1}^\infty\int_0^\infty t^{r-1}\mathsf{P}\left\{\sup_{(2^{n-1}-1)t^\gamma < u^\gamma \leq (2^n-1)t^\gamma}\left[|X_u| - \varepsilon u^\gamma\right] \geq \varepsilon t^\gamma\right\}\,\mathrm{d}t\\
&\leq r\sum_{n=1}^\infty\int_0^\infty t^{r-1}\mathsf{P}\left\{\sup_{u^\gamma\leq 2^n t^\gamma}|X_u| \geq 2^{n-1}\varepsilon t^\gamma\right\}\,\mathrm{d}t\\
&= r\sum_{n=1}^\infty\int_0^\infty t^{r-1}\mathsf{P}\left\{M_{2^{n/\gamma}u} \geq 2^{n-1}\varepsilon t^\gamma\right\}\,\mathrm{d}t\\
&= r\left[\sum_{n=1}^\infty 2^{-n/\gamma}\right]\int_0^\infty u^{r-1}\mathsf{P}\left\{M_u \geq (\varepsilon/2)u^\gamma\right\}\,\mathrm{d}u.
\end{aligned}$$

It follows that

$$\mathsf{E}\left[L_{2\varepsilon}(\gamma)^r\right] \leq r\left(2^{1/\gamma}-1\right)^{-1}\int_0^\infty u^{r-1}\mathsf{P}\left\{M_u \geq (\varepsilon/2)u^\gamma\right\}\,\mathrm{d}u \leq \tag{16}$$

$$\leq r\left(2^{1/\gamma}-1\right)^{-1}\sum_{n=1}^\infty n^{r-1}\mathsf{P}\left\{\max_{1\leq t\leq n} X_n \geq \varepsilon n^\gamma\right\} \tag{17}$$

which yields the implication (15) and completes the proof. □

The following theorem shows that, in the i.i.d. case, the implications in (13) become equivalences.

Theorem 3. *Let $\{S_n\}_{n\in\mathbb{Z}_+}$ be the random walk with mean $\mathsf{E}[S_n] = \mu n$. The following statements are equivalent*

$$\mathsf{E}|S_1|^{r+1} < \infty \Longleftrightarrow n^{-1}S_n \xrightarrow[n\to\infty]{r-completely} \mu, \tag{18}$$

$$\mathsf{E}|S_1|^{r+1} < \infty \Longleftrightarrow n^{-1}S_n \xrightarrow[n\to\infty]{r-quickly} \mu, \tag{19}$$

$$\mathsf{E}|S_1|^{r+1} \Longleftrightarrow \sum_{n=1}^\infty n^{r-1}\mathsf{P}\left\{\sup_{k\geq n}\frac{1}{k}\,|S_k - \mu| > \varepsilon\right\} < \infty \quad \textit{for all } \varepsilon > 0. \tag{20}$$

Proof. By Theorem 1, in the i.i.d. case,

$$\mathsf{E}|S_1|^{r+1} < \infty \Longleftrightarrow \sum_{n=1}^\infty n^{r-1}\mathsf{P}\left(\frac{1}{n}\,|S_n - \mu| > \varepsilon\right) < \infty \quad \forall\varepsilon > 0 \tag{21}$$

and

$$\mathsf{E}|S_1|^{r+1} < \infty \iff \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(\sup_{k\geq n}\frac{1}{k}\,|S_k-\mu| > \varepsilon\right) < \infty \quad \forall\varepsilon > 0, \tag{22}$$

so that assertion (18) follows from (21) and (20) from (22).

Next, let

$$L_\varepsilon = \sup\left\{n \geq 1 : |S_n - n\,\mu| \geq n\,\varepsilon\right\} \quad (\sup\varnothing = 0).$$

By Lemma 2(i),

$$\mathsf{E}[L_\varepsilon^r] \leq r\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left\{\sup_{t\geq n}\left(|S_t - \mu\,t|/t\right) \geq \varepsilon\right\} \quad \forall\varepsilon > 0, \tag{23}$$

which along with (22) implies (19). □

2.4. Further Remarks on r-Complete Convergence, r-Quick Convergence and Rates of Convergence in SLLN

Let $\{S_n\}_{n\in\mathbb{Z}_+}$ be a random walk. Without loss of generality let $S_0 = 0$ and $\mathsf{E}[S_1] = 0$.

1. Strassen [3] proved, in particular, that if $f(n) = (2n\log n)^{1/2}$ in Lemma 2, then for $r > 0$

$$\limsup_{n\to\infty}\frac{S_n}{\sqrt{2n\log n}} = \sqrt{r\,\mathsf{E}[S_1^2]} \quad r-\text{quickly} \tag{24}$$

whenever $\mathsf{E}|S_1|^p < \infty$ for $p > (2r+1)$. He also proved the functional form of the law of the iterated logarithm.

2. Lai [7] improved this result showing that Strassen's moment condition $\mathsf{E}|S_1|^p < \infty$ for $p > (2r+1)$ can be relaxed. Specifically, he showed that a weaker condition

$$\mathsf{E}\left[|S_1|^{2(r+1)}(\log^+|S_1|+1)^{-(r+1)})\right] < \infty \quad \text{for } r > 0 \tag{25}$$

is the best one can do (i.e., both necessary and sufficient):

$$\mathsf{E}\left[|S_1|^{2(r+1)}(\log^+|S_1|+1)^{-(r+1)}\right] < \infty \iff \limsup_{n\to\infty}\frac{S_n}{\sqrt{2n\log n}} < \infty \;\; r-\text{quickly},$$

in which case equality (24) holds.

Note, however, that for $r = 0$ in terms of the a.s. convergence

$$\mathsf{E}\left[|S_1|^2\right] < \infty \iff \limsup_{n\to\infty}\frac{S_n}{\sqrt{2n\log\log n}} = \sqrt{\mathsf{E}\left[|S_1|^2\right]} \;\; \text{a.s.}$$

but under condition (25) for all $r > 0$

$$\limsup_{n\to\infty}\frac{S_n}{\sqrt{2n\log\log n}} = \infty \;\; r-\text{quickly}.$$

3. Let $\alpha > 1/2$ and $r > 0$. Chow and Lai [9] established the following one-sided inequality for tail probabilities:

$$\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(\max_{1\leq t\leq n} S_t \geq n^\alpha\right) \leq C_{r,\alpha}\left\{\mathsf{E}\left[(S_1^+)^{(r+1)/\alpha}\right] + \left(\mathsf{E}[S_1^2]\right)^{r/(2\alpha-1)}\right\} \tag{26}$$

whenever $\mathsf{E}|S_1|^2 < \infty$. Under the same hypotheses, this one-sided inequality implies the two-sided one:

$$\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(\max_{1\leq t\leq n} |S_t| \geq n^\alpha\right) \leq C_{r,\alpha}\left\{\mathsf{E}\left[|S_1|^{(r+1)/\alpha}\right] + \left(\mathsf{E}[S_1^2]\right)^{r/(2\alpha-1)}\right\}. \tag{27}$$

The upper bound in (27) turns out to be sharp since the lower bound also holds:

$$\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(\max_{1\le t\le n}|S_t|\ge n^{\alpha}\right)\ge 1+B_{r,\alpha}\left\{\mathsf{E}\left[|S_1|^{(r+1)/\alpha}\right]+\left(\mathsf{E}[S_1^2]\right)^{r/(2\alpha-1)}\right\}.$$

Here the constants $C_{r,\alpha}$ and $B_{r,\alpha}$ are universal depending only on r,α.

The results of Chow and Lai [9] provide one-sided analogues of the results of Baum and Katz [2] as well as extend their results. Indeed, the one-sided inequality (26) implies that the following statements are equivalent for the zero-mean random walk:

(i) $\mathsf{E}[(S_1^+)^{(r+1)/\alpha}]<\infty$;

(ii) $\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(n^{-\alpha}S_n\ge\varepsilon\right)<\infty$ for all $\varepsilon>0$;

(iii) $\sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left(\sup_{k\ge n}k^{-\alpha}S_k\ge\varepsilon\right)<\infty$ for all $\varepsilon>0$,

where $\alpha>1/2$.

Clearly, the two-sided inequality (27) yields the assertions of Theorem 1 if $\mu=0$.

4. The Marcinkiewicz-Zygmund SLLN states that for $\alpha>1/2$ the following implications hold:

$$\mathsf{E}|S_1|^{1/\alpha}<\infty \Longleftrightarrow n^{-\alpha}S_n \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{28}$$

The strengthened r-quick equivalent of this SLLN is: For any $r>0$ and $\alpha>1/2$ the following statements are equivalent,

$$\begin{aligned}\mathsf{E}[|S_1|^{(r+1)/\alpha}]<\infty &\Longleftrightarrow \sum_{i=1}^{\infty} n^{r-1}\mathsf{P}\left\{\frac{1}{n^{\alpha}}|S_n|>\varepsilon\right\}<\infty \ \text{ for all } \varepsilon>0\\ &\Longleftrightarrow \sum_{n=1}^{\infty} n^{r-1}\mathsf{P}\left\{\sup_{k\ge n}\frac{1}{k^{\alpha}}|S_k|>\varepsilon\right\}<\infty \ \text{ for all } \varepsilon>0\\ &\Longleftrightarrow n^{-\alpha}S_n \xrightarrow[n\to\infty]{r-\text{quickly}} 0.\end{aligned} \tag{29}$$

Implications (29) follow from Theorem 1, Theorem 3 and inequality (27). The proof is almost obvious and omitted.

3. Applications of r-Complete and r-Quick Convergences in Statistics

In this section, we outline certain statistical applications which show the usefulness of r-complete and r-quick versions of the SLLN.

3.1. Sequential Hypothesis Testing

We begin with formulating the following multihypothesis testing problem for a general non-i.i.d stochastic model. Let $(\Omega,\mathscr{F},\mathscr{F}_n,\mathsf{P})$, $n\in\mathbb{Z}_+=\{0,1,2,\dots\}$, be a filtered probability space with standard assumptions about the monotonicity of the sub-σ-algebras $\mathscr{F}_n$. The sub-σ-algebra $\mathscr{F}_n=\sigma(\mathbf{X}^n)$ of $\mathscr{F}$ is assumed to be generated by the sequence $\mathbf{X}^n=\{X_t,\,1\le t\le n\}$ observed up to time n, which is defined on the space $(\Omega,\mathscr{F})$. The hypotheses are $\mathsf{H}_i:\ \mathsf{P}=\mathsf{P}_i$, $i=0,1,\dots,N$, where $\mathsf{P}_0,\mathsf{P}_1,\dots,\mathsf{P}_N$ are given probability measures assumed to be locally mutually absolutely continuous, i.e., their restrictions $\mathsf{P}_i^{\{n\}}$ and $\mathsf{P}_j^{\{n\}}$ to $\mathscr{F}_n$ are equivalent for all $1\le n<\infty$ and all $i,j=0,1,\dots,N$, $i\ne j$. Let $\mathsf{Q}^{\{n\}}$ be a restriction to $\mathscr{F}_n$ of a σ-finite measure Q on $(\Omega,\mathscr{F})$. Under P_i the sample $\mathbf{X}^n=(X_1,\dots,X_n)$ has a joint density $p_{i,n}(\mathbf{X}^n)$ with respect to the dominating measure $\mathsf{Q}^{(n)}$ for all $n\in\mathbb{Z}_+$, which can be written as

$$p_{i,n}(\mathbf{X}^n)=\prod_{t=1}^{n} f_{i,t}(X_t|\mathbf{X}^{t-1}), \tag{30}$$

where $f_{i,n}(X_n|\mathbf{X}^{n-1})$, $n\ge 1$ are corresponding conditional densities.

Define the likelihood ratio (LR) process between the hypotheses H_i and H_j

$$\Lambda_{ij}(n) = \frac{\mathrm{d}\mathsf{P}_i^{\{n\}}}{\mathrm{d}\mathsf{P}_j^{\{n\}}}(\mathbf{X}^n) = \frac{p_{i,n}(\mathbf{X}^n)}{p_{j,n}(\mathbf{X}^n)} = \prod_{t=1}^{n} \frac{f_{i,t}(X_t|\mathbf{X}^{t-1})}{f_{j,t}(X_t|\mathbf{X}^{t-1})}$$

and the log-likelihood ratio (LLR) process

$$\lambda_{ij}(n) = \log \Lambda_{ij}(n) = \sum_{t=1}^{n} \log\left[\frac{f_{i,t}(X_t|\mathbf{X}^{t-1})}{f_{j,t}(X_t|\mathbf{X}^{t-1})}\right],$$

where we set $\Lambda_{ij}(0) = 1$ and $\lambda_{ij}(0) = 0$.

A multihypothesis sequential test is a pair $\delta = (d, T)$, 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 $\{0, 1, \dots, N\}$. Specifically, $d = i$ means that the hypothesis H_i is accepted upon stopping, i.e., $\{d = i\} = \{T < \infty, \delta \text{ accepts } \mathsf{H}_i\}$. Let $\alpha_{ij}(\delta) = \mathsf{P}_i(d = j)$, $i \neq j$, $i, j = 0, 1, \dots, N$, denote the error probabilities of the test δ, i.e., the probabilities of accepting the hypothesis H_j when H_i is true.

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

$$\mathbb{C}(\boldsymbol{\alpha}) = \left\{\delta : \alpha_{ij}(\delta) \le \alpha_{ij} \text{ for } i, j = 0, 1, \dots, N,\ i \neq j\right\}, \tag{31}$$

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

Let E_i denote the expectation under the hypothesis H_i (i.e., under the measure P_i). The goal of a statistician is to find a sequential test that would minimize the expected sample sizes $\mathsf{E}_i[T]$ for all hypotheses H_i, $i = 0, 1, \dots, N$ at least approximately, say asymptotically for small probabilities of errors, i.e., as $\alpha_{ij} \to 0$.

3.1.1. Asymptotic Optimality of Walds's SPRT

Assume first that $N = 1$, i.e., that we are dealing with two hypotheses H_0 and H_1. In the mid 1940s, Wald [11,12] introduced the *Sequential Probability Ratio Test* (SPRT) for the sequence of i.i.d. observations $X_1, X_2, \dots$, in which case $f_{i,t}(X_t|\mathbf{X}^{t-1}) = f_i(X_t)$ in (30) and the LR $\Lambda_{1,0}(n) = \Lambda_n$ is

$$\Lambda_n = \prod_{t=1}^{n} \frac{f_1(X_t)}{f_0(X_t)}.$$

After n observations have been made Wald's SPRT prescribes for each $n \ge 1$:

$$\begin{aligned} &\text{Stop and accept } H_1 &&\text{if} \quad \Lambda_n \ge A_1. \\ &\text{Stop and accept } H_0 &&\text{if} \quad \Lambda_n \le A_0. \\ &\text{Continue sampling} &&\text{if} \quad A_0 < \Lambda_n < A_1. \end{aligned}$$

where $A_0 < 1 < A_1$ are two thresholds.

Let $Z_t = \log[f_1(X_t)/f_0(X_t)]$ be the LLR for the observation X_t, so the LLR for the sample $\mathbf{X}^n$ is the sum

$$\lambda_{10}(n) = \lambda_n = \sum_{t=1}^{n} Z_t, \quad n = 1, 2, \dots$$

Let $a_0 = -\log A_0 < 0$ and $a_1 = \log A_1 > 0$. The SPRT $\delta_*(a_0, a_1) = (d_*, T_*)$ can be represented in the form

$$T_*(a_0, a_1) = \inf\{n \ge 1 : \lambda_n \notin (-a_0, a_1)\}, \quad d_*(a_0, a_1) = \begin{cases} 1 & \text{if } \lambda_{T_*} \ge a_1 \\ 0 & \text{if } \lambda_{T_*} \le -a_0. \end{cases} \tag{32}$$

In the case of two hypotheses, the class of tests (31) is of the form

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

i.e., it upper-bounds the probabilities of errors of Type 1 (false positive) $\alpha_0(\delta) = \alpha_{0,1}(\delta)$ and Type 2 (false negative) $\alpha_1(\delta) = \alpha_{1,0}(\delta)$, respectively.

Wald's SPRT has an extraordinary optimality property: it minimizes both expected sample sizes $\mathsf{E}_0[T]$ and $\mathsf{E}_1[T]$ in the class of sequential (and non-sequential) tests $\mathbb{C}(\alpha_0, \alpha_1)$ with given error probabilities as long as the observations are i.i.d. under both hypotheses. More specifically, Wald and Wolfowitz [13] proved, using a Bayesian approach, that if $\alpha_0 + \alpha_1 < 1$ and thresholds $-a_0$ and a_1 can be selected in such a way that $\alpha_0(\delta_*) = \alpha_0$ and $\alpha_1(\delta_*) = \alpha_1$, then the SPRT δ_* is strictly optimal in class $\mathbb{C}(\alpha_0, \alpha_1)$. A rigorous proof of this fundamental result is tedious and involves several delicate technical details. Alternative proofs can be found in [14–19].

Regardless of the strict optimality of SPRT which holds if, and only if, thresholds are selected so that the probabilities of errors of SPRT are exactly equal to the prescribed values α_0, α_1, which is usually impossible, suppose that thresholds a_0 and a_1 are so selected that

$$a_0 \sim \log(1/\alpha_1) \quad \text{and} \quad a_1 \sim \log(1/\alpha_0) \quad \text{as } \alpha_{\max} \to 0. \tag{33}$$

Then

$$\mathsf{E}_1[T_*] \sim \frac{|\log \alpha_0|}{I_1}, \quad \mathsf{E}_0[T_*] \sim \frac{|\log \alpha_1|}{I_0} \quad \text{as } \alpha_{\max} \to 0, \tag{34}$$

where $I_1 = \mathsf{E}_1[Z_1]$ and $I_0 = \mathsf{E}_0[-Z_1]$ are Kullback-Leibler (K-L) information numbers so that the following asymptotic lower bounds for ESS are attained by SPRT:

$$\inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_1[T] \ge \frac{|\log \alpha_0|}{I_1} + o(1), \quad \inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_0[T] \ge \frac{|\log \alpha_1|}{I_0} + o(1) \quad \text{as } \alpha_{\max} \to 0$$

(cf. [6]). Hereafter $\alpha_{\max} = \max(\alpha_0, \alpha_1)$. The following inequalities for the error probabilities of the SPRT hold in the most general non-i.i.d. case

$$\alpha_1(\delta_*) \le \exp\{-a_0\}[1 - \alpha_0(\delta_*)], \quad \alpha_0(\delta_*) \le \exp\{-a_1\}[1 - \alpha_1(\delta_*)]. \tag{35}$$

These bounds can be used to guarantee asymptotic relations (33).

In the i.i.d. case, by the SLLN, the LLR λ_n has the following stability property

$$n^{-1}\lambda_n \xrightarrow[n\to\infty]{\mathsf{P}_1-\text{a.s.}} I_1, \quad n^{-1}(-\lambda_n) \xrightarrow[n\to\infty]{\mathsf{P}_0-\text{a.s.}} I_0. \tag{36}$$

This allows one to conjecture that if in the general non-i.i.d. case the LLR is also stable in the sense that the almost sure convergence conditions (36) are satisfied with some positive and finite numbers I_1 and I_0, then the asymptotic formulas (34) still hold. In the general case, these numbers represent the local K–L information in the sense that often (while not always) $I_1 = \lim_{n\to\infty} n^{-1}\mathsf{E}_1[\lambda_n]$ and $I_0 = \lim_{n\to\infty} n^{-1}\mathsf{E}_0[-\lambda_n]$. Note, however, that in the general non-i.i.d. case the SLLN does not even guarantee the finiteness of the expected sample sizes $\mathsf{E}_i[T_*]$ of the SPRT, so some additional conditions are needed, such as a certain rate of convergence in the strong law, e.g., complete or quick convergence.

In 1981, Lai [8] was the first who proved asymptotic optimality of Wald's SPRT in a general non-i.i.d. case as $\alpha_{\max} = \max(\alpha_0, \alpha_1) \to 0$. While the motivation was near optimality of invariant SPRTs with respect to nuisance parameters, Lai proved a more general result using the r-quick convergence concept. Specifically, for $i = 0, 1$ and $0 < I_i < \infty$, define

$$L_1(\varepsilon) = \sup\{n \ge 1 : |\lambda_n - I_1| \ge \varepsilon\} \quad \text{and} \quad L_0(\varepsilon) = \sup\{n \ge 1 : |\lambda_n + I_0| \ge \varepsilon\}$$

($\sup\{\varnothing\} = 0$) and suppose that $\mathsf{E}_i[L_i(\varepsilon)^r] < \infty$ for some $r > 0$ and every $\varepsilon > 0$, i.e., that the normalized LLR converges r-quickly to I_1 under P_1 and to $-I_0$ under P_0:

$$n^{-1}\lambda_n \xrightarrow[n\to\infty]{\mathsf{P}_1-r-\text{quickly}} I_1 \quad \text{and} \quad n^{-1}\lambda_n \xrightarrow[n\to\infty]{\mathsf{P}_0-r-\text{quickly}} -I_0. \tag{37}$$

Strengthening the a.s. convergence (36) into the r-quick version (37), Lai [8] established first-order asymptotic optimality of Wald's SPRT for moments of the stopping time distribution up to order r: If thresholds $a_i(\alpha_0, \alpha_1)$, $i = 0, 1$ in the SPRT are so selected that $\delta_*(a_0, a_1) \in \mathbb{C}(\alpha_0, \alpha_1)$ and asymptotics (33) hold, then as $\alpha_{\max} \to 0$,

$$\begin{aligned} \inf_{\delta\in\mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_1[T^r] &\sim \left(\frac{|\log\alpha_0|}{I_1}\right)^r \sim \mathsf{E}_1[T_*^r], \\ \inf_{\delta\in\mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_0[T^r] &\sim \left(\frac{|\log\alpha_1|}{I_0}\right)^r \sim \mathsf{E}_0[T_*^r]. \end{aligned} \tag{38}$$

Wald's ideas have been generalized in many publications to construct sequential tests of composite hypotheses with nuisance parameters when these hypotheses can be reduced to simple ones by the principle of invariance. If M_n is the maximal invariant statistic and $p_i(M_n)$ is the density of this statistic under hypothesis H_i, then the invariant SPRT is defined as in (32) with the LLR $\lambda_n = \log[p_1(M_n)/p_0(M_n)]$. But even if the observations $X_1, X_2, \dots$ are i.i.d. the invariant LLR statistic λ_n is not a random walk anymore and Wald's methods cannot be applied directly. Lai [8] has applied the asymptotic optimality property (38) of Wald's SPRT in the non-i.i.d. case to investigate optimality properties of several classical invariant SPRTs such as the sequential t-test, the sequential T^2-test, and Savage's rank-order test.

In the sequel, the case where the a.s. convergence in the non-i.i.d. model (36) holds with the rate $1/n$ we will call *asymptotically stationary*. Assume now that (36) is generalized to

$$\lambda_n/\psi(n) \xrightarrow[n\to\infty]{\mathsf{P}_1-\text{a.s.}} I_1, \quad (-\lambda_n)/\psi(n) \xrightarrow[n\to\infty]{\mathsf{P}_0-\text{a.s.}} I_0, \tag{39}$$

where $\psi(t)$ is a positive increasing function. If $\psi(t)$ is not linear, then this case will be referred to as the *asymptotically non-stationary*. A simple example where this generalization is needed is testing H_0 versus H_1 regarding the mean of the normal distribution:

$$X_n = i\, S_n + \xi_n, \quad n \in \mathbb{Z}_+, \quad i = 0, 1,$$

where $\{\xi_n\}_{n\ge1}$ is a zero-mean i.i.d. standard Gaussian sequence $\mathcal{N}(0,1)$ and $S_n = \sum_{j=0}^k c_j n^j$ is a polynomial of order $k > 1$. Then

$$\lambda_n = \sum_{t=1}^n S_t X_t - \frac{1}{2}\sum_{t=1}^n S_t^2,$$

$\mathsf{E}_1[\lambda_n] = -\mathsf{E}_0[\lambda_n] = \frac{1}{2}\sum_{t=1}^n S_t^2 \sim c_k^2 n^{2k}$ for large n, so $\psi(n) = n^{2k}$ and $I_1 = I_0 = c_k^2/2$ in (39). This example is of interest for certain practical applications, in particular, for the recognition of ballistic objects and satellites [19].

Tartakovsky et al. [6, Sec 3.4] generalized Lai's results for the asymptotically non-stationary case. Write $\Psi(t)$ for the inverse function for $\psi(t)$.

Theorem 4. *Assume that there exist finite positive numbers I_0 and I_1 and an increasing nonnegative function $\psi(t)$ such that the r-quick convergence conditions*

$$\frac{\lambda_n}{\psi(n)} \xrightarrow[n\to\infty]{\mathsf{P}_1-r-\text{quickly}} I_1, \quad \frac{-\lambda_n}{\psi(n)} \xrightarrow[n\to\infty]{\mathsf{P}_0-r-\text{quickly}} I_0$$

hold. If thresholds $a_0(\alpha_0, \alpha_1)$ and $a_1(\alpha_0, \alpha_1)$ are selected so that $\delta_*(a_0, a_1) \in \mathbb{C}(\alpha_0, \alpha_1)$ and $a_0 \sim |\log \alpha_1|$ and $a_1 \sim |\log \alpha_0|$, then, as $\alpha_{\max} \to 0$,

$$\begin{aligned} \inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_1[T^r] &\sim \left[\Psi\left(\frac{|\log \alpha_0|}{I_1}\right)\right]^r \sim \mathsf{E}_1[T_*^r], \\ \inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_0[T^r] &\sim \left[\Psi\left(\frac{|\log \alpha_1|}{I_0}\right)\right]^r \sim \mathsf{E}_0[T_*^r]. \end{aligned} \tag{40}$$

This theorem implies that the SPRT asymptotically minimizes the moments of the stopping time distribution up to the order r.

The proof of this theorem is performed in two steps which are related to our previous discussion of the rates of convergence in Section 2. The first step is to obtain the asymptotic lower bounds in class $\mathbb{C}(\alpha_0, \alpha_1)$:

$$\liminf_{\alpha_{\max}\to 0} \frac{\inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_1[T^r]}{[\Psi(|\log \alpha_0|/I_1)]^r} \ge 1, \quad \liminf_{\alpha_{\max}\to 0} \frac{\inf_{\delta \in \mathbb{C}(\alpha_0,\alpha_1)} \mathsf{E}_0[T^r]}{[\Psi(|\log \alpha_1|/I_0)]^r} \ge 1.$$

These bounds hold whenever the following right-tail conditions for the LLR are satisfied:

$$\begin{aligned} \lim_{M\to\infty} \mathsf{P}_1\left\{\frac{1}{\psi(M)} \max_{1\le n\le M} \lambda_n \ge (1+\varepsilon) I_1\right\} = 1, \\ \lim_{M\to\infty} \mathsf{P}_0\left\{\frac{1}{\psi(M)} \max_{1\le n\le M} (-\lambda_n) \ge (1+\varepsilon) I_0\right\} = 1. \end{aligned}$$

Note that by Lemma 1 these conditions are satisfied when the SLLN (39) holds so that the almost sure convergence (39) is sufficient. However, as we already mentioned, the SLLN for the LLR is not sufficient to guarantee even the finiteness of the SPRT stopping time.

The second step is to show that the lower bounds are attained by the SPRT. To do so, it suffices to impose the following additional left-tail conditions:

$$\sum_{n=1}^{\infty} n^{r-1} \mathsf{P}_1\left\{\lambda_n \le (I_1 - \varepsilon)\psi(n)\right\} < \infty, \quad \sum_{n=1}^{\infty} n^{r-1} \mathsf{P}_0\left\{-\lambda_n \le (I_0 - \varepsilon)\psi(n)\right\} < \infty$$

for all $0 < \varepsilon < \min(I_0, I_1)$. Since both right-tail and left-tail conditions hold if the LLR converges r-completely to I_i,

$$\sum_{n=1}^{\infty} n^{r-1} \mathsf{P}_1\left\{\left|\frac{\lambda_n}{\psi(n)} - I_1\right| \ge \varepsilon\right\} < \infty, \quad \sum_{n=1}^{\infty} n^{r-1} \mathsf{P}_0\left\{\left|\frac{\lambda_n}{\psi(n)} + I_0\right| \ge \varepsilon\right\}$$

and since r-quick convergence implies r-complete convergence (see (13)), we conclude that the assertions (40) hold.

Remark 4. *In the i.i.d. case, Wald's approach allows us to establish asymptotic equalities (40) with $I_1 = \mathsf{E}_1[\lambda_1]$ and $I_0 = -\mathsf{E}_0[\lambda_1]$ being K-L information numbers under the only condition of finiteness I_i. However, Wald's approach breaks down in the non-i.i.d. case. Certain generalizations in the case of independent but non-identically and substantially non-stationary observations, extending Wald's ideas, have been considered in [19–22]. Theorem 4 covers all these non-stationary models.*

Fellouris and Tartakovsky [23] extended previous results on asymptotic optimality of the SPRT to the case of multistream hypothesis testing problem when the observations are sequentially acquired in multiple data streams (or channels or sources). The problem is to test the null hypothesis H_0 that none of the N streams is affected against the composite hypothesis H_B that a subset $B \subset \{1, \dots, N\}$ is affected. Two sequential tests were studied in [23] – the Generalized Sequential Likelihood Ratio Test and the Mixture Sequential Likelihood Ratio Test. It has been shown that both tests are first-order asymptotically optimal, minimizing moments of the sample size $\mathsf{E}_0[T^r]$ and $\mathsf{E}_B[T^r]$ for all $B \in \mathscr{P}$ up to order r as $\max(\alpha_0, \alpha_1) \to 0$ in the class of tests

$$\mathbb{C}_{\mathscr{P}}(\alpha_0, \alpha_1) = \left\{ \delta : \mathsf{P}_0(d=1) \le \alpha_0 \ \text{ and } \ \max_{B \in \mathscr{P}} \mathsf{P}_B(d=0) \le \alpha_1 \right\}, \quad 0 < \alpha_i < 1,$$

where P_B is the distribution of observations under hypothesis H_B and $\mathscr{P}$ is a class of subsets of $\{1, \dots, N\}$ that incorporates prior information which is available regarding the subset of affected streams, e.g., not more than $K < N$ streams can be affected.[1] The proof is essentially based on the concept of r-complete convergence of LLR with the rate $1/n$. See also Chapter 1 in [5].

3.1.2. Asymptotic Optimality of the Multihypothesis SPRT

We now return to the multihypothesis model with $N > 1$ that we started to discuss at the beginning of this section (see (30) and (31)). The problem of sequential testing of many hypotheses is substantially more difficult than that of testing two hypotheses. For multiple-decision testing problems, it is usually very difficult, if even possible, to obtain optimal solutions. Finding an optimal non-Bayesian test in the class of tests (31) that minimizes ESS $\mathsf{E}_i[T]$ for all hypotheses H_i, $i = 0, 1, \dots, N$ is not manageable even in the i.i.d. case. For this reason, a substantial part of the development of sequential multihypothesis testing in the 20th century has been directed towards the study of certain combinations of one-sided sequential probability ratio tests when observations are i.i.d. (see, e.g., [24–29]).

We will focus on the following first-order asymptotic criterion: Find a multihypothesis test $\delta_*(\boldsymbol{\alpha}) = (d_*(\boldsymbol{\alpha}), T_*(\boldsymbol{\alpha}))$ such that for some $r > 0$

$$\lim_{\alpha_{\max} \to 0} \frac{\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{E}_i[T^r]}{\mathsf{E}_i[T_*(\boldsymbol{\alpha})^r]} = 1 \quad \text{for all } i = 0, 1, \dots, N, \tag{41}$$

where $\alpha_{\max} = \max_{0 \le i,j \le N, i \ne j} \alpha_{ij}$.

In 1998, Tartakovsky [4] was the first who considered the sequential multiple hypothesis testing problems for general non-i.i.d. stochastic models following Lai's idea of exploiting the r-quick convergence in the SLLN for two hypotheses. The results have been obtained for both discrete and continuous-time scenarios and for the asymptotically non-stationary case where the LLR processes between hypotheses converge to finite numbers with the rate $1/\psi(t)$. Two multihypothesis tests were investigated: (1) The *Rejecting* test which rejects the hypotheses one by one and the last hypothesis, which is not rejected, is accepted, and (2) The *Matrix Accepting* test that accepts a hypothesis for which all component SPRTs that involve this hypothesis vote for accepting it. We now proceed with introducing this accepting test which we will refer to as the *Matrix SPRT* (MSPRT). In the present article, we do not consider the continuous-time scenarios. Those who are interested in continuous time we refer to [4,6,20,22,30].

Write $\mathcal{N} = \{0, 1, \dots, N\}$. For a threshold matrix $(A_{ij})_{i,j \in \mathcal{N}}$, with $A_{ij} > 0$ and the A_{ii} are immaterial (say 0), define the Matrix SPRT $\delta_*^N = (T_*^N, d_*^N)$, built on $(N+1)N/2$ one-sided SPRTs between the hypotheses H_i and H_j, as follows:

$$\text{Stop at the first } n \ge 1 \text{ such that, for some } i,\ \Lambda_{ij}(n) \ge A_{ji} \text{ for all } j \ne i, \tag{42}$$

and accept the unique H_i that satisfies these inequalities. Note that for $N = 1$ the MSPRT coincides with Wald's SPRT.

[1] In many practical problems, K is substantially smaller than the total number of streams N, which can be very large.

In the following, we omit the superscript N in $\delta_*^N = (T_*^N, d_*^N)$ for brevity. Obviously, with $a_{ji} = \log A_{ji}$, the MSPRT in (42) can be written as

$$\begin{aligned} & T_* = \inf\left\{n \ge 1 : \lambda_{ij}(n) \ge a_{ji} \ \text{ for all } j \ne i \text{ and some } i\right\}, && (43)\\ & d_* = i \ \text{ for which (43) holds.} && (44)\end{aligned}$$

Introducing the Markov accepting times for the hypotheses H_i as

$$T_i = \inf\left\{n \ge 1 : \lambda_{i0}(n) \ge \max_{\substack{1 \le j \le N\\ j \ne i}} [\lambda_{j0}(n) + a_{ji}]\right\}, \quad i = 0, 1, \dots, N, \tag{45}$$

the test in (43)–(44) can be also written in the following form:

$$T_* = \min_{0 \le j \le N} T_j, \qquad d_* = i \quad \text{if} \quad T_* = T_i. \tag{46}$$

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

Using Wald's likelihood ratio identity, it is easily shown that $\alpha_{ij}(\delta_*) \le \exp(-a_{ij})$ for $i, j \in \mathcal{N}$, $i \ne j$, so selecting $a_{ji} = |\log \alpha_{ji}|$ implies $\delta_* \in \mathbb{C}(\boldsymbol{\alpha})$. These inequalities are similar to Wald's ones in the binary hypothesis case and are very imprecise. In his ingenious paper, Lorden [28] showed that with a very sophisticated design that includes accurate estimation of thresholds accounting for overshoots, the MSPRT is nearly optimal in the third-order sense, i.e., it minimizes ESS for all hypotheses up to an additive disappearing term: $\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{E}_i[T] = \mathsf{E}_i[T_*] + o(1)$ as $\alpha_{\max} \to 0$. This result holds only for i.i.d. models with the finite second moment $\mathsf{E}_i[\lambda_{ij}(1)^2] < \infty$. In non-i.i.d. cases (and even for i.i.d. for higher moments $r > 1$), there is no way to obtain such a result, so we focus on the first-order optimality (41).

The following theorem establishes asymptotic operating characteristics and optimality of MSPRT under the r-quick convergence of $\lambda_{ij}(n)/\psi(n)$ to finite K-L-type numbers I_{ij}, where $\psi(n)$ is a positive increasing function, $\psi(\infty) = \infty$.

Theorem 5 (MSPRT asymptotic optimality)**.** *Assume that there exist finite positive numbers I_{ij}, $i, j = 0, 1, \dots, N$, $i \ne j$ and an increasing nonnegative function $\psi(t)$ such that for some $r > 0$*

$$\frac{\lambda_{ij}(n)}{\psi(n)} \xrightarrow[n \to \infty]{\mathsf{P}_i - r - quickly} I_{ij} \quad \textit{for all } \ i, j = 0, 1, \dots, N, i \ne j. \tag{47}$$

Then the following assertions are true.

(i) *For $i = 0, 1, \dots, N$,*

$$\mathsf{E}_i[T_*^r] \sim \left[\Psi\left(\max_{\substack{0 \le j \le N\\ j \ne i}} \frac{a_{ji}}{I_{ij}}\right)\right]^r \quad \textit{as} \quad \min_{j,i} a_{ji} \to \infty. \tag{48}$$

(ii) *If the thresholds are so selected that $\alpha_{ij}(\delta^*) \le \alpha_{ij}$ and $a_{ji} \sim |\log \alpha_{ji}|$, in particular as $a_{ji} = |\log \alpha_{ji}|$, then for all $i = 0, 1, \dots, N$*

$$\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{E}_i[T^r] \sim \left[\Psi\left(\max_{\substack{0 \le j \le N\\ j \ne i}} \frac{|\log \alpha_{ji}|}{I_{ij}}\right)\right]^r \sim \mathsf{E}_i[T_*^r] \quad \textit{as } \alpha_{\max} \to 0. \tag{49}$$

Assertion (ii) implies that the MSPRT minimizes asymptotically the moments of the stopping time distribution up to order r for all hypotheses $\mathsf{H}_0, \mathsf{H}_1, \dots, \mathsf{H}_N$ in the class of tests $\mathbb{C}(\boldsymbol{\alpha})$.

Remark 5. *Both assertions of Theorem 5 are correct under the r-complete convergence*

$$\frac{\lambda_{ij}(n)}{\psi(n)} \xrightarrow[n\to\infty]{\mathsf{P}_i-r-complete} I_{ij} \quad \text{for all} \;\; i,j = 0,1,\dots,N, i \neq j,$$

i.e., whenever

$$\sum_{n=1}^{\infty} n^{r-1} \mathsf{P}_i \left\{ \frac{1}{\psi(n)} \left|\lambda_{ij}(n) - I_{ij}\right| > \varepsilon \right\} < \infty \quad \text{for all } \varepsilon > 0.$$

While this statement was not proved anywhere so far, it can be easily proved using the methods developed for multistream hypothesis testing and changepoint detection [5, Ch 1, Ch 6].

Remark 6. *As the example given in Subsection 3.4.3 of [6] shows, the r-quick convergence conditions in Theorem 5 (or corresponding r-complete convergence conditions for LLR processes) cannot be generally relaxed into the almost sure convergence*

$$\frac{\lambda_{ij}(n)}{\psi(n)} \xrightarrow[n\to\infty]{\mathsf{P}_i-a.s.} I_{ij} \quad \text{for all} \;\; i,j = 0,1,\dots,N, i \neq j. \tag{50}$$

However, the following weak asymptotic optimality result holds for the MSPRT under the a.s. convergence: if the a.s. convergence (50) holds with the power function $\psi(t) = t^k$, $k > 0$, then for every $0 < \varepsilon < 1$,

$$\inf_{\delta \in \mathbb{C}(\boldsymbol{\alpha})} \mathsf{P}_i\left(T > \varepsilon\, T_*\right) \to 1 \quad \text{as} \;\; \alpha_{\max} \to 0 \;\; \text{for all} \;\; i = 0,1,\dots,N \tag{51}$$

whenever thresholds a_{ji} are selected as in Theorem 5(ii).

Note that several interesting statistical and practical applications of these results to invariant sequential testing and multisample slippage scenarios are discussed in Sections 4.5 and 4.6 of Tartakovsky et al. [6] (see Mosteller [31] and Ferguson [16] for terminology regarding multisample slippage problems).

3.2. Sequential Changepoint Detection

Sequential changepoint detection (or quickest disorder detection) is an important branch of Sequential Analysis. In the sequential setting, one assumes that the observations are made successively, one at a time, and as long as their behavior suggests that the process of interest is in a normal state, the process is allowed to continue; if the state is believed to have become anomalous, the goal is to detect the change in distribution as rapidly as possible. Quickest change detection problems have an enormous number of important applications, e.g., object detection in noise and clutter, industrial quality control, environment surveillance, failure detection, navigation, seismology, computer network security, genomics, epidemiology (see, e.g., [32–41]). Several challenging application areas are discussed in the books by Tartakovsky, Nikiforov, and Basseville [6, Ch 11] and Tartakovsky [5, Ch 8].

3.2.1. Changepoint Models

The probability distribution of the observations $\mathbf{X} = \{X_n\}_{n\in\mathbb{Z}_+}$, which are acquired sequentially in time, is subject to a change at an unknown point in time $\nu \in \{0,1,2,\dots\}$, so that $X_1,\dots,X_\nu$ are generated by one stochastic model and $X_{\nu+1}, X_{\nu+2},\dots$ by another model. A sequential detection rule is a stopping time T for an observed sequence $\{X_n\}_{n\ge1}$, i.e., T is an integer-valued random variable, such that the event $\{T = n\}$ belongs to the sigma-algebra $\mathscr{F}_n = \sigma(X_1,\dots,X_n)$ generated by observations $X_1,\dots,X_n$.

Let P_∞ denote the probability measure corresponding to the sequence of observations $\{\mathbf{X}_n\}_{n\ge1}$ when there is never a change ($\nu = \infty$) and, for $k = 0,1,\dots$, let P_k denote the measure corresponding to the sequence $\{\mathbf{X}_n\}_{n\ge 1}$ when $\nu = k < \infty$. By $\mathsf{H}_\infty : \nu = \infty$ we denote the hypothesis that the change never occurs and by $\mathsf{H}_k : \nu = k$ – the hypothesis that the change occurs at time $0 \le k < \infty$.

Consider first a general non-i.i.d. model assuming that the observations may have a very general stochastic structure. Specifically, if we let as before $\mathbf{X}^n = (X_1, \dots, X_n)$ denote the sample of size n, then when $\nu = \infty$ (there is no change) the conditional density of X_n given $\mathbf{X}^{n-1}$ is $g_n(X_n|\mathbf{X}^{n-1})$ for all $n \ge 1$ and when $\nu = k < \infty$, then the conditional density of X_n given $\mathbf{X}^{n-1}$ is $g_n(X_n|\mathbf{X}^{n-1})$ for $n \le k$ and $f_n(X_n|\mathbf{X}^{n-1})$ for $n > k$. Thus, for the general non-i.i.d. changepoint model, the joint density $p(\mathbf{X}^n|\mathsf{H}_k)$ under hypothesis H_k can be written as follows

$$p(\mathbf{X}^n|\mathsf{H}_k) = \begin{cases} \prod_{t=1}^n g_t X_t|\mathbf{X}^{t-1}) & \text{for } \ \nu = k \ge n, \\ \prod_{t=1}^k g_t(\mathbf{X}_t|\mathbf{X}^{t-1}) \times \prod_{t=k+1}^n f_t(X_t|\mathbf{X}^{t-1}) & \text{for } \ \nu = k < n, \end{cases} \tag{52}$$

where $g_n(X_n|\mathbf{X}^{n-1})$ is the pre-change conditional density and $f_n(X_n|\mathbf{X}^{n-1})$ is the post-change conditional density which may depend on ν, $f_n(X_n|\mathbf{X}^{n-1}) = f_n^{(\nu)}(X_n|\mathbf{X}^{n-1})$, but we will omit the superscript ν for brevity.

The classical changepoint detection problem deals with the i.i.d. case where there is a sequence of observations $X_1, X_2, \dots$ that are identically distributed with a probability density function (pdf) $g(x)$ for $n \le \nu$ and with a pdf $f(x)$ for $n > \nu$. That is, in the i.i.d. case, the joint density of the vector $\mathbf{X}^n = (X_1, \dots, X_n)$ under hypothesis H_k in (52) is simplified as

$$p(\mathbf{X}^n|\mathsf{H}_k) = \begin{cases} \prod_{t=1}^n g(X_t) & \text{for } \nu = k \ge n, \\ \prod_{t=1}^k g(X_t) \times \prod_{t=k+1}^n f(X_t) & \text{for } \nu = k < n. \end{cases} \tag{53}$$

Note that, as discussed in [5,6], in applications, there are two different kinds of changes – additive and non-additive. Additive changes lead to a change in the mean value of the sequence of observations. Non-additive changes are typically produced by a change in variance or covariance, i.e., these are spectral changes.

We now proceed with discussing the models for the change point ν. The change point ν may be considered either as an unknown deterministic number or as a random variable. If the change point is treated as a random variable, then the model has to be supplied with the *prior distribution* of the change point. There may be several changepoint mechanisms and, as a result, a random variable ν may be partially or completely dependent on the observations or independent of the observations. To account for these possibilities at once, let $\pi_{-1} = \Pr(\nu < 0)$ and $\pi_k = \Pr(\nu = k|\mathbf{X}^k)$, $k \ge 0$, and observe that π_k, $k = 1, 2, \dots$ are $\mathscr{F}_k$-adapted. That is, the probability of a change occurring at the time instant $\nu = k$ depends on $\mathbf{X}^k$, the observations' history accumulated up to and including the time $k \ge 1$. The probability $\pi_{-1} + \pi_0 = \Pr(\nu \le 0)$ represents the probability of the "atom" associated with the event that the change already took place before the observations became available. With the so-defined prior distribution, one can describe very general changepoint models, including those that assume ν to be a $\{\mathscr{F}_n\}$-adapted stopping time (see Moustakides [42]). In this article, we will not discuss Moustakides's concept by allowing the prior distribution to depend on some additional information available to "Nature" (see [5] for a detailed discussion); rather when considering a Bayesian approach we will assume that the prior distribution of the unknown change point is independent of the observations.

3.2.2. Popular Changepoint Detection Procedures

Before formulating criteria of optimality in the next subsection, we begin with defining the three most popular and common change detection procedures, which are either optimal or nearly optimal in different settings. To define these procedures we need to introduce the partial likelihood ratio and the corresponding log-likelihood ratio

$$\mathsf{LR}_t = \frac{f_t(X_t|\mathbf{X}^{t-1})}{g_t(X_t|\mathbf{X}^{t-1})}, \quad Z_t = \log\frac{f_t(X_t|\mathbf{X}^{t-1})}{g_t(X_t|\mathbf{X}^{t-1})}, \quad t = 1, 2, \dots$$

It is worth iterating that for general non-i.i.d. models the post-change density often depends on the point of change, $f_t(X_t|\mathbf{X}^{t-1}) = f_t^{(\nu)}(X_t|\mathbf{X}^{t-1})$, so in general $\mathsf{LR}_t = \mathsf{LR}_t^{(\nu)}$ and $Z_t = Z_t^{(\nu)}$ also depend on the change point ν. However, this is not the case for the i.i.d. model (53).

The CUSUM Procedure

We now introduce the *Cumulative Sum* (CUSUM) algorithm, which was first proposed by Page [43] for the i.i.d. model (53). Recall that we consider the changepoint detection problem as a problem of testing two hypotheses: H_ν that the change occurs at a fixed point $0 \le \nu < \infty$ against the alternative H_∞ that the change never occurs. The LR between these hypotheses is $\Lambda_n^\nu = \prod_{t=\nu+1}^n \mathsf{LR}_t$ for $\nu < n$ and 1 for $\nu \ge n$. Since the hypothesis H_ν is composite, we may apply the generalized likelihood ratio (GLR) approach maximizing the LR Λ_n^ν over ν to obtain the GLR statistic

$$V_n = \max_{0 \le \nu < n} \prod_{t=\nu+1}^{n} \mathsf{LR}_t, \quad n \ge 1.$$

It is easy to verify that this statistic follows the recursion

$$V_n = \max\{1, V_{n-1}\}\mathsf{LR}_n, \quad n \ge 1, \ \ V_0 = 1 \tag{54}$$

as long as the partial LR LR_n does not depend on the change point, i.e., the post-change conditional density $f_n(X_n|\mathbf{X}^{n-1})$ does not depend on ν. This is always the case for i.i.d. models (53) when $f_n(X_n|\mathbf{X}^{n-1}) = f(X_n)$. However, as we already mentioned, for non-i.i.d. models often $f_n(X_n|\mathbf{X}^{n-1}) = f_n^{(\nu)}(X_n|\mathbf{X}^{n-1})$ depends on the change point ν, so $\mathsf{LR}_n = \mathsf{LR}_n^{(\nu)}$, in which case recursion (54) does not hold.

The logarithmic version of V_n, $W_n = \log V_n$, is related to Page's CUSUM statistic G_n introduced by Page [43] in the i.i.d. case as $G_n = \max(0, W_n)$. In fact, the statistic G_n can also be obtained via the GLR approach by maximizing the LLR $\lambda_n^\nu = \log \Lambda_n^\nu$ over $0 \le \nu < \infty$. However, since the hypotheses H_∞ and H_ν are indistinguishable for $\nu \ge n$ the maximization over $\nu \ge n$ does not make too much sense. Note also that in contrast to Page's CUSUM statistic G_n the statistic W_n may take values smaller than 0, so the CUSUM procedure

$$\mathsf{T}_{\mathsf{CS}} = \inf\{n \ge 1 : W_n \ge a\} \tag{55}$$

makes sense even for negative values of the threshold a. Thus, it is more general than Page's CUSUM. Note the recursions

$$W_n = W_{n-1}^+ + Z_n, \quad n \ge 1, \ \ W_0 = 0 \tag{56}$$

and

$$G_n = (G_{n-1} + Z_n)^+, \quad n \ge 1, \ \ G_0 = 0$$

in case where $Z_n = \log[f_n(X_n|\mathbf{X}^{n-1})/g_n(X_n|\mathbf{X}^{n-1})]$ does not depend on ν.

Shiryaev's Procedure

In the i.i.d. case and for the zero-modified geometric prior distribution of the change point, Shiryaev [44] introduced the change detection procedure that prescribes thresholding of the posterior probability $\mathsf{P}(\nu < n|\mathbf{X}^n)$. Introducing the statistic

$$\mathsf{S}_n^\pi = \frac{\mathsf{P}(\nu < n|\mathbf{X}^n)}{1-\mathsf{P}(\nu < n|\mathbf{X}^n)}$$

one can write the stopping time of the Shiryaev procedure in the general non-i.i.d. case and for an arbitrary prior π as

$$\mathsf{T}_{\mathsf{SH}} = \inf\{n \ge 1 : \mathsf{S}_n^\pi \ge A\}, \tag{57}$$

where A is a threshold controlling for the false alarm risk. Write $\pi_{-1} = \mathsf{P}(\nu < 0) = p$, $p \in [0,1)$. The statistic S_n^π can be written as

$$\begin{aligned} \mathsf{S}_n^\pi &= \frac{p}{1-p}\Lambda_n^0 + \frac{1}{\mathsf{P}(\nu \ge n)}\sum_{k=0}^{n-1}\pi_k\Lambda_n^k \\ &= \frac{p}{1-p}\prod_{t=1}^{n}\mathsf{LR}_t + \frac{1}{\mathsf{P}(\nu \ge n)}\sum_{k=0}^{n-1}\pi_k\prod_{t=k+1}^{n}\mathsf{LR}_t, \quad n \ge 1, \quad \mathsf{S}_0^\pi = \frac{p}{1-p}, \end{aligned} \tag{58}$$

where the product $\prod_{t=i}^{j}\mathsf{LR}_t = 1$ for $j < i$. Threshold A has to be set larger than $p/(1-p)$ to avoid triviality, since otherwise $\mathsf{T}_{\mathsf{SH}} = 0$ w.p. 1.

Often (following Shiryaev's assumptions) it is supposed that the change point ν is distributed according to the zero-modified geometric distribution Geometric(p,ϱ)

$$\mathsf{P}(\nu < 0) = \pi_{-1} = p \;\text{ and }\; \mathsf{P}(\nu = k) = (1-p)\varrho(1-\varrho)^k \quad \text{for } k = 0,1,2,\dots, \tag{59}$$

where $p \in [0,1)$ and $\varrho \in (0,1)$.

If LR_n does not depend on the change point ν and the prior distribution is zero-modified geometric (59) then the statistic $\widetilde{\mathsf{S}}_n^\varrho = \mathsf{S}_n^\pi/\varrho$ can be rewritten in the recursive form

$$\widetilde{\mathsf{S}}_n^\varrho = \left(1 + \widetilde{\mathsf{S}}_{n-1}^\varrho\right)\frac{\mathsf{LR}_n}{1-\varrho}, \quad n \ge 1, \quad \widetilde{\mathsf{S}}_0^\varrho = \frac{p}{(1-p)\varrho}. \tag{60}$$

However, as mentioned above, this may not be the case for non-i.i.d. models since often LR_n depends on ν.

Shiryaev–Roberts Procedure

The generalized Shiryaev–Roberts (SR) change detection procedure is based on thresholding of the generalized SR statistic

$$\mathsf{R}_n^{r_0} = r_0\Lambda_n^0 + \sum_{k=0}^{n-1}\Lambda_n^k = r_0\prod_{t=1}^{n}\mathsf{LR}_t + \sum_{k=0}^{n-1}\prod_{t=k+1}^{n}\mathsf{LR}_t, \quad n \ge 1, \tag{61}$$

with a non-negative head-start $\mathsf{R}_0 = r_0$, $r_0 \ge 0$, i.e., the stopping time of the SR procedure is given by

$$\mathsf{T}_{\mathsf{SR}}^{r_0} = \inf\left\{n \ge 1 : \mathsf{R}_n^{r_0} \ge A\right\}, \quad A > 0. \tag{62}$$

This procedure is usually referred to as the SR-r detection procedure in contrast to the standard SR procedure $\mathsf{T}_{\mathsf{SR}} \equiv \mathsf{T}_{\mathsf{SR}}^{r_0}, r_0 = 0$ that starts with a zero initial condition $r_0 = 0$. In the i.i.d. case (53), this modification of the SR procedure was introduced and studied in detail in [45,46].

If LR_n does not depend on the change point ν, then the SR-r detection statistic satisfies the recursion

$$\mathsf{R}_n^{r_0} = (1 + \mathsf{R}_{n-1}^{r_0})\mathsf{LR}_n, \quad n \geq 1, \quad \mathsf{R}_0^{r_0} = r_0.$$

Note that as the parameter of the geometric prior distribution $\varrho \to 0$, the Shiryaev statistic $\widetilde{\mathsf{S}}_n^{\varrho}$ converges to the SR-r statistic $\mathsf{R}_n^{r_0}$.

3.2.3. Optimality Criteria

The goal of online change detection is to detect the change as soon as possible after it occurs controlling a false alarm rate at a given level. Tartakovsky et al. [6, Sec. 6.3] suggested five changepoint problem settings – the Bayesian approach, the generalized Bayesian approach, the minimax approach, the uniform (pointwise) approach, and the approach related to multicyclic detection of a change in a stationary regime. In this article, we discuss only a single-run case and two main settings – Bayesian and uniform pointwise optimality, which are tightly related.

Let E_k denote the expectation with respect to the measure P_k when the change occurs at $\nu = k < \infty$ and E_∞ with respect to P_∞ when there is no change.

In 1954, Page [43] suggested measuring the risk associated with 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. He called these performance characteristics the *Average Run Length* (ARL). Page also introduced the now most famous change detection procedure – CUSUM procedure – and analyzed it using these operating characteristics.

While the false alarm rate is reasonable to measure by the ARL to false alarm

$$\mathsf{ARL2FA}(T) = \mathsf{E}_\infty[T],$$

as Figure 1 suggests, the risk associated with a true change detection is reasonable to measure by the conditional average delay to detection

$$\mathsf{CEDD}_\nu(T) = \mathsf{E}_\nu[T - \nu | T > \nu], \quad \nu = 0, 1, \dots,$$

but not necessarily by the ARL to detection $\mathsf{E}_0[T] \equiv \mathsf{CEDD}_0(T)$. A good detection procedure should guarantee small values of the expected detection delay $\mathsf{CEDD}_\nu(T)$ for all change points $\nu \geq 0$ when $\mathsf{ARL2FA}(T)$ is fixed at a certain level. However, if the false alarm risk is measured in terms of the ARL to false alarm, i.e., it is required that $\mathsf{ARL2FA}(T) \geq \gamma$ for some $\gamma \geq 1$, then a procedure that minimizes the conditional average delay to detection $\mathsf{CEDD}_\nu(T)$ uniformly over all ν does not exist. For this reason, we have to resort to different optimality criteria, e.g., to Bayesian and minimax criteria.

doi:10.20944/preprints202305.0904.v1

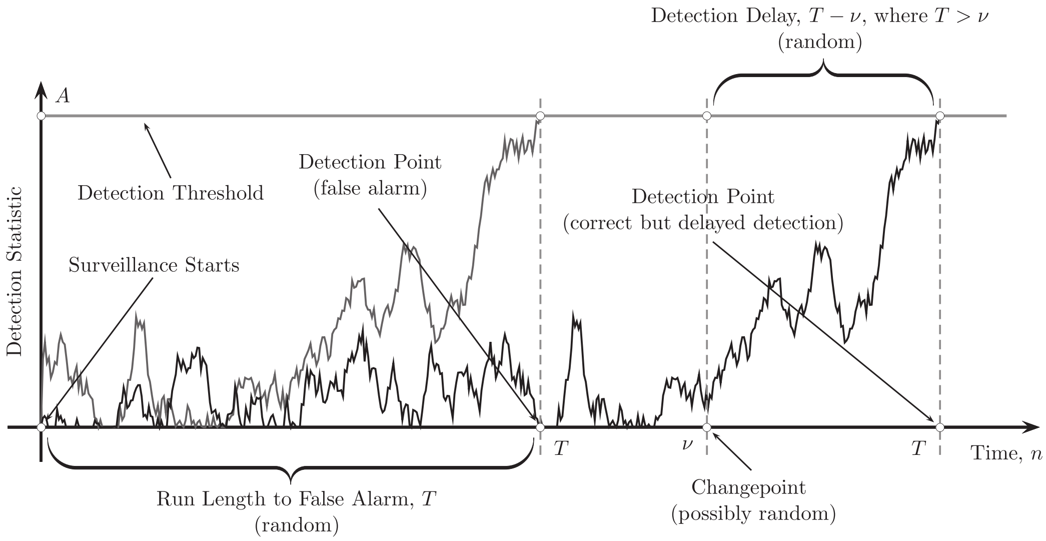

Figure 1. Illustration of single-run sequential changepoint detection. Two possibilities in the detection process: false alarm (left) and correct detection (right).

Minimax Changepoint Optimization Criteria

There are two popular minimax criteria. The first one was introduced by Lorden [47]:

$$\inf_{T}\sup_{\nu\ge 0}\operatorname{ess\,sup}\mathsf{E}_\nu[T-\nu\mid T>\nu,\mathscr{F}_\nu]\quad\text{subject to }\mathsf{ARL2FA}(T)\ge\gamma.$$

It requires minimizing the conditional expected delay to detection $\mathsf{E}_\nu[T-\nu\mid T>\nu,\mathscr{F}_\nu]$ in the worst-case scenario with respect to both the change point ν and the trajectory $(X_1,\dots,X_\nu)$ of the observed process in the class of detection procedures

$$\mathbb{C}_{\mathrm{ARL}}(\gamma)=\{T:\mathsf{ARL2FA}(T)\ge\gamma\},\quad\gamma\ge 1,$$

for which the ARL to false alarm exceeds the prespecified value $\gamma\in[1,\infty)$. Let $\mathsf{ESADD}(T)=\sup_{\nu\ge 0}\operatorname{ess\,sup}\mathsf{E}_\nu[T-\nu\mid T>\nu,\mathscr{F}_\nu]$ denote Lorden's speed detection measure. Under Lorden's minimax approach the goal is to find a stopping time $T_{\mathrm{opt}}\in\mathbb{C}_{\mathrm{ARL}}(\gamma)$ such that

$$\mathsf{ESADD}(T_{\mathrm{opt}})=\inf_{T\in\mathbb{C}_{\mathrm{ARL}}(\gamma)}\mathsf{ESADD}(T)\quad\text{for any }\gamma\ge 1.$$

In the classical i.i.d. scenario (53), Lorden [47] proved that the CUSUM detection procedure (55) is asymptotically first-order minimax optimal as $\gamma\to\infty$, i.e.,

$$\inf_{T\in\mathbb{C}_{\mathrm{ARL}}(\gamma)}\mathsf{ESADD}(T)=\mathsf{ESADD}(\mathsf{T}_{\mathsf{CS}})(1+o(1)),\quad\gamma\to\infty.$$

Later on, Moustakides [48], using optimal stopping theory, in his ingenious paper established the exact optimality of CUSUM for any ARL to false alarm $\gamma\ge 1$.

Another popular, less pessimistic minimax criterion is due to Pollak [49]:

$$\inf_{T}\sup_{\nu\ge 0}\mathsf{CEDD}_\nu(T)\quad\text{subject to }\mathsf{ARL2FA}(T)\ge\gamma,$$

which requires minimizing the conditional expected delay to detection $\mathsf{CEDD}_\nu(T) = \mathsf{E}_\nu[T - \nu \mid T > \nu]$ in the worst-case scenario with respect to the change point ν in class $\mathbb{C}_{\mathrm{ARL}}(\gamma)$. Under Pollak's minimax approach the goal is to find a stopping time $T_{\mathrm{opt}} \in \mathbb{C}_{\mathrm{ARL}}(\gamma)$ such that

$$\sup_{\nu \geq 0} \mathsf{CEDD}_\nu(T_{\mathrm{opt}}) = \inf_{T \in \mathbb{C}_{\mathrm{ARL}}(\gamma)} \sup_{\nu \geq 0} \mathsf{CEDD}_\nu(T) \quad \text{for any } \gamma \geq 1.$$

For the i.i.d. model (53), Pollak [49] showed that the modified SR detection procedure that starts from the quasi-stationary distribution of the SR statistic (i.e., the head-start r_0 in the SR-r procedure is a specific random variable) is third-order asymptotically optimal as $\gamma \to \infty$, i.e., the best one can attain up to an additive term $o(1)$:

$$\inf_{T \in \mathbb{C}_{\mathrm{ARL}}(\gamma)} \sup_{\nu \geq 0} \mathsf{CEDD}_\nu(T) = \sup_{\nu \geq 0} \mathsf{CEDD}_\nu(\mathsf{T}_{\mathsf{SR}}^{\mathsf{r}_0}) + o(1), \quad \gamma \to \infty,$$

where $o(1) \to 0$ as $\gamma \to \infty$. Later Tartakovsky et al. [50] proved that this is also true for the SR-r procedure (62) that starts from the fixed but specially designed point $r_0 = r_0(\gamma)$ that depends on γ, which was first introduced and thoroughly studied by Moustakides et al. [45]. See also Polunchenko and Tartakovsky [51] on the exact optimality of the SR-r procedure.

Bayesian Changepoint Optimization Criterion

In Bayesian problems, the point of change ν is treated as random with a prior distribution $\pi_k = \Pr(\nu = k)$, $-\infty < k < +\infty$. Define the probability measure on the Borel σ-algebra $\mathscr{B}$ in $\mathbb{R}^\infty \times \mathbb{N}$ as

$$\mathsf{P}^\pi(\mathcal{A} \times \mathcal{K}) = \sum_{k \in \mathcal{K}} \pi_k \mathsf{P}_k(\mathcal{A}), \ \ \mathcal{A} \in \mathscr{B}(\mathbb{R}^\infty), \ \ \mathcal{K} \in \mathbb{N}.$$

Under measure P^π the change point ν has distribution $\pi = \{\pi_k\}$ and the model for the observations is given in (52). From the Bayesian point of view, it is reasonable to measure the false alarm risk with the *Weighted Probability of False Alarm* (PFA), defined as

$$\mathsf{PFA}^\pi(T) := \mathsf{P}^\pi(T \leq \nu) = \sum_{k=-\infty}^{\infty} \pi_k \mathsf{P}_k(T \leq k) = \sum_{k=0}^{\infty} \pi_k \mathsf{P}_\infty(T \leq k). \tag{63}$$

The summation in (63) is over $k \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ since $\mathsf{P}_\infty(T < 0) = 0$. Also, the last equality follows from the fact that $\mathsf{P}_k(T \leq k) = \mathsf{P}_\infty(T \leq k)$ because the event $\{T \leq k\}$ depends on the first k observations which under measure P_k correspond to the no-change hypothesis H_∞. Thus, for $\alpha \in (0, 1)$, introduce the class of changepoint detection procedures

$$\mathbb{C}_\pi(\alpha) = \{T : \mathsf{PFA}^\pi(T) \leq \alpha\} \tag{64}$$

for which the weighted PFA does not exceed a prescribed level α. Let E^π denote expectation with respect to measure P^π.

Shiryaev [18,44] introduced the Bayesian optimality criterion

$$\inf_{T \in \mathbb{C}_\pi(\alpha)} \mathsf{E}^\pi[(T - \nu)^+],$$

which is equivalent to minimizing the conditional average detection delay $\mathsf{EDD}^\pi(T) = \mathsf{E}^\pi[T - \nu | T > \nu]$

$$\inf_T \mathsf{EDD}^\pi(T) \quad \text{subject to } \mathsf{PFA}^\pi(T) \leq \alpha.$$

Under the Bayesian approach, the goal is to find a stopping time $T_{\rm opt} \in \mathbb{C}_\pi(\alpha)$ such that

$$\mathsf{EDD}^\pi(T_{\rm opt}) = \inf_{T\in\mathbb{C}_\pi(\alpha)} \mathsf{EDD}^\pi(T) \quad \text{for any } \alpha \in (0,1). \tag{65}$$

For the i.i.d. model (53) and under the assumption that the changepoint ν has the zero-modified geometric prior distribution Geometric(p, ϱ) (59), this problem was solved by Shiryaev [18,44]. Shiryaev [18,44,52] proved that the optimal detection procedure is based on comparing the posterior probability of a change currently being in effect with a certain detection threshold, which is equivalent to the stopping time $\mathsf{T}_{\mathsf{SH}}(A)$ (57). To guarantee its strict optimality the detection threshold $A = A_\alpha$ should be set to guarantee that the PFA is exactly equal to the selected level α. Thus, if $A = A_\alpha$ can be selected in such a way that $\mathsf{PFA}^\pi(\mathsf{T}_{\mathsf{SH}}(A_\alpha)) = \alpha$, then it is strictly optimal in class $\mathbb{C}_\pi(\alpha)$,

$$\inf_{T\in\mathbb{C}_\pi(\alpha)} \mathsf{EDD}^\pi(T) = \mathsf{EDD}^\pi(\mathsf{T}_{\mathsf{SH}}(A_\alpha)) \ \text{ for any } \ 0 < \alpha < 1 - p.$$

Uniform Optimality Under Local Probabilities of False Alarm

While the Bayesian and minimax formulations are reasonable and can be justified in many applications, it would be most desirable to guarantee small values of the conditional expected detection delay $\mathsf{CEDD}_\nu(T) = \mathsf{E}_\nu[T - \nu | T \ge \nu]$ uniformly for all $\nu \in \mathbb{Z}_+$ when the false alarm risk is fixed at a certain level. However, as we already mentioned, if the false alarm risk is measured in terms of the ARL to false alarm, i.e. if it is required that $\mathsf{ARL2FA}(T) \ge \gamma$ for some $\gamma \ge 1$, then a procedure that minimizes $\mathsf{CEDD}_\nu(T)$ for all ν does not exist. More importantly, as discussed in [5, Sec 2.3], the requirement of having large values of the $\mathsf{ARL2FA}(T)$ generally does not guarantee small values of the maximal local probability of false alarm $\mathsf{MLPFA}(T) = \sup_{\ell\ge 0} \mathsf{P}_\infty(T \le \ell + m | T > \ell)$ in a time window of a length $m \ge 1$, while the opposite is always true (see Lemmas 2.1-2.2 in [5]). Hence, the constraint $\mathsf{MLPFA}(T) \le \beta$ is more stringent than $\mathsf{ARL2FA}(T) \ge \gamma$.

Yet another reason for considering the MLPFA constraint instead of the ARL to false alarm constraint is that the latter one makes sense, if and only if, the P_∞-distribution of stopping times is geometric or at least close to geometric, which is often the case for many popular detection procedures such as CUSUM and SR in the i.i.d. case. However, for general non-i.i.d. models this is not necessarily true (see [5, Sec 2.3] and [53] for a detailed discussion).

For these reasons, introduce the most stringent class of change detection procedures for which the $\mathsf{MLPFA}(T)$ is upper-bounded by the prespecified level $\beta \in (0,1)$:

$$\mathbb{C}_{\rm PFA}(m,\beta) = \left\{ T : \sup_{\ell\ge 0} \mathsf{P}_\infty(T \le \ell + m | T > \ell) \le \beta \right\}. \tag{66}$$

The goal is to find a stopping time $T_{\rm opt} \in \mathbb{C}_{\rm PFA}(m,\beta)$ such that

$$\mathsf{CEDD}_\nu(T_{\rm opt}) = \inf_{T\in\mathbb{C}_{\rm PFA}(m,\beta)} \mathsf{CEDD}_\nu(T) \quad \text{for all } \nu \in \mathbb{Z}_+ \text{ and any } 0 < \beta < 1. \tag{67}$$

3.2.4. Asymptotic Optimality for General Non-i.i.d. Models via r-Quick and r-Complete Convergence

Complete Convergence and General Bayesian Changepoint Detection Theory

Consider first the Bayesian problem assuming that the change point ν is a random variable independent of the observations with a prior distribution $\pi = \{\pi_k\}$. Unfortunately, in the general non-i.i.d. case and for an arbitrary prior π, the Bayesian optimization problem (65) is intractable for arbitrary values of PFA $\alpha \in (0,1)$. For this reason, we will consider the following first-order asymptotic problem assuming that the given PFA α approaches zero: Find a change detection procedure T^* such that it minimizes the expected detection delay $\mathsf{EDD}^{\pi}(T)$ asymptotically to first order as $\alpha \to 0$. That is, the goal is to design such a detection procedure T^* that

$$\inf_{T \in \mathbb{C}_{\pi}(\alpha)} \mathsf{EDD}^{\pi}(T) = \mathsf{EDD}^{\pi}(T^*)(1+o(1)) \quad \text{as } \alpha \to 0, \tag{68}$$

where $o(1) \to 0$ as $\alpha \to 0$. It turns out that in the asymptotic setting, it is also possible to find a procedure that minimizes the conditional expected detection delay $\mathsf{EDD}_k(T) = \mathsf{E}_k\left[T-k \,|\, T>k\right]$ uniformly for all possible values of the change point $\nu = k \in \mathbb{Z}_+$, i.e.,

$$\lim_{\alpha \to 0} \frac{\inf_{T \in \mathbb{C}_{\pi}(\alpha)} \mathsf{EDD}_k(T)}{\mathsf{EDD}_k(T^*)} = 1 \quad \text{for all } k \in \mathbb{Z}_+. \tag{69}$$

Note that if the change occurs before the observations become available, i.e., $\nu = k \in \{-1, -2, \dots\}$, then $\mathsf{EDD}_k(T) \equiv \mathsf{E}_0[T]$ since $T \geq 0$ w.p. 1.

Furthermore, asymptotic optimality results can be also established for higher moments of the detection delay of order $r > 1$

$$\mathsf{E}_k\left[(T-k)^r \,|\, T>k\right] \quad \text{and} \quad \mathsf{E}^{\pi}\left[(T-\nu)^r \,|\, T>\nu\right].$$

Since the Shiryaev procedure $T_{\mathsf{SH}}(A)$ defined in (57)-(58) is optimal for the i.i.d. model and Geometric(p,ϱ) prior, it is reasonable to assume that it is asymptotically optimal for the more general prior and the non-i.i.d model. However, to study asymptotic optimality we need certain constraints imposed on the prior distribution and on the asymptotic behavior of the decision statistics as the sample size increases, i.e., on the general stochastic model (52).

Assume that the prior distribution $\{\pi_k\}$ is fully supported, i.e., $\pi_k > 0$ for all $k \in \mathbb{Z}_+$ and $\pi_\infty = 0$ and that the following conditions hold:

$$\lim_{n \to \infty} \frac{1}{n} \left| \log \sum_{k=n+1}^{\infty} \pi_k \right| = \mu \quad \text{for some } 0 \leq \mu < \infty; \tag{70}$$

$$\sum_{k=0}^{\infty} \pi_k |\log \pi_k|^r < \infty \quad \text{for some } r \geq 1 \quad \text{if } \mu = 0. \tag{71}$$

Note that if $\mu > 0$, then by condition (70) the prior distribution has an exponential right tail. Distributions such as geometric and discrete versions of gamma and logistic distributions, i.e., models with bounded hazard rates, belong to this class. In this case, condition (71) holds automatically. If $\mu = 0$, the distribution has a heavy tail, i.e., belongs to the model with a vanishing hazard rate. However, we cannot allow this distribution to have a too-heavy tail, which is guaranteed by condition (71).

Define the LLR of the hypotheses H_k and H_∞

$$\lambda_n^k = \log \frac{d\mathsf{P}_k^{(n)}}{d\mathsf{P}_\infty^{(n)}} = \sum_{t=k+1}^{n} \frac{f_t(X_t|\mathbf{X}^t)}{g_t(X_t|\mathbf{X}^t)}, \quad n > k$$

($\lambda_n^k = 0$ for $n \leq k$). To obtain asymptotic optimality results the general non-i.i.d. model for observations is restricted to the case that the normalized LLR $n^{-1}\lambda_{k+n}^k$ obeys the SLLN as $n \to \infty$ with a finite and positive number I under the probability measure P_k and its r-complete strengthened version

$$\sum_{n=1}^{\infty} n^{r-1} \sup_{k \in \mathbb{Z}_+} \mathsf{P}_k\left\{|n^{-1}\lambda_{k+n}^k - I| > \varepsilon\right\} < \infty \quad \text{for every } \varepsilon > 0. \tag{72}$$

By Lemma 7.2.1 in [6],

$$\mathsf{PFA}^{\pi}(\mathsf{T}_{\mathsf{SH}}(A)) \le 1/(1+A) \quad \text{for every } A > p/(1-p),$$

and therefore, setting $A = A_\alpha = (1-\alpha)/\alpha$ guarantees that $\mathsf{T}_{\mathsf{SH}}(A_\alpha) \in \mathbb{C}_\pi(\alpha)$.

The following theorem that can be deduced from Theorem 3.7 in [5] shows that the Shiryaev detection procedure is asymptotically optimal if the normalized LLR $n^{-1}\lambda_{k+n}^{k}$ converges r-completely to a positive and finite number I and the prior distribution satisfies conditions (70)-(71).

Theorem 6. *Let $r \ge 1$. Let the prior distribution of the change point satisfy conditions (70)-(71). Assume that there exists some number $0 < I < \infty$ such that the LLR process $n^{-1}\lambda_{k+n}^{k}$ converges to I uniformly completely as $n \to \infty$ under P_k, i.e., condition (72) holds. If threshold $A = A_\alpha$ in the Shiryaev procedure is so selected that $\mathsf{PFA}^{\pi}(\mathsf{T}_{\mathsf{SH}}(A_\alpha)) \le \alpha$ and $\log A_\alpha \sim |\log \alpha|$ as $\alpha \to 0$, e.g., as $A = (1-\alpha)/\alpha$, then as $\alpha \to 0$*

$$\inf_{T \in \mathbb{C}_\pi(\alpha)} \mathsf{E}_k\left[(T-k)^r \mid T > k\right] \sim \left(\frac{|\log \alpha|}{I+\mu}\right)^r \sim \mathsf{E}_k\left[(\mathsf{T}_{\mathsf{SH}}-k)^r \mid \mathsf{T}_{\mathsf{SH}} > k\right] \quad \textit{for all } k \in \mathbb{Z}_+$$

and

$$\inf_{T \in \mathbb{C}_\pi(\alpha)} \mathsf{E}^{\pi}\left[(T-\nu)^r \mid T > \nu\right] \sim \left(\frac{|\log \alpha|}{I+\mu}\right)^r \sim \mathsf{E}^{\pi}\left[(\mathsf{T}_{\mathsf{SH}}-\nu)^r \mid \mathsf{T}_{\mathsf{SH}} > \nu\right].$$

Therefore, the Shiryaev procedure $\mathsf{T}_{\mathsf{SH}}(A_\alpha)$ is first-order asymptotically optimal as $\alpha \to 0$ in class $\mathbb{C}_\pi(\alpha)$, minimizing moments of the detection delay up to order r whenever the r-complete version of the SLLN (72) holds for the LLR process.

For $r = 1$, the assertions of this theorem imply asymptotic optimality of the Shiryaev procedure for the expected detection delays (68) and (69) as well as asymptotic approximations for the expected detection delays.

Remark 7. *The results of Theorem 6 can be generalized to the asymptotically non-stationary case where $\lambda_{k+n}^{k}/\psi(n)$ converges to I uniformly completely as $n \to \infty$ under P_k with a non-linear function $\psi(n)$ similarly to the hypothesis testing problem discussed in Section 3.1. See also the recent paper [54] for the minimax change detection problem with independent but substantially non-stationary post-change observations.*

It is also interesting to see how two other most popular changepoint detection procedures – the SR and CUSUM – perform in the Bayesian context.

Consider the SR-r procedure defined by (61)-(62). It follows from Lemma 3.4 (page 100) in [5] that

$$\mathsf{PFA}^{\pi}(\mathsf{T}_{\mathsf{SR}}^{r_0}(A)) \le \frac{r_0 \sum_{k=1}^{\infty} \pi_k + \sum_{k=1}^{\infty} k\pi_k}{A} \quad \text{for every } A > 0,$$

and therefore, setting $A = A_\alpha = \alpha^{-1}(r_0 + \sum_{k=1}^{\infty} k\pi_k)$ implies $\mathsf{T}_{\mathsf{SR}}^{r_0}(A_\alpha) \in \mathbb{C}_\pi(\alpha)$. Let threshold $A = A_\alpha$ in the SR-r procedure is so selected that $\mathsf{PFA}_{\pi}(\mathsf{T}_{\mathsf{SR}}^{r_0}(A_\alpha)) \le \alpha$ and $\log A_\alpha \sim |\log \alpha|$ as $\alpha \to 0$, e.g., as $A_\alpha = \alpha^{-1}(r_0 + \sum_{k=1}^{\infty} k\pi_k)$, then as $\alpha \to 0$

$$\mathsf{E}_k\left[(\mathsf{T}_{\mathsf{SR}}^{r_0}-k)^r \mid \mathsf{T}_{\mathsf{SR}}^{r_0} > k\right] \sim \left(\frac{|\log \alpha|}{I}\right)^r \quad \text{for all } k \in \mathbb{Z}_+ \tag{73}$$

and

$$\mathsf{E}^{\pi}\left[(\mathsf{T}_{\mathsf{SR}}^{r_0}-\nu)^r \mid \mathsf{T}_{\mathsf{SR}}^{r_0} > \nu\right] \sim \left(\frac{|\log \alpha|}{I}\right)^r \tag{74}$$

whenever the uniform r-complete convergence condition (72) holds. Therefore, the SR-r procedure $\mathsf{T}_{\mathsf{SR}}^{r_0}(A_\alpha)$ is first-order asymptotically optimal as $\alpha \to 0$ in class $\mathbb{C}_\pi(\alpha)$, minimizing moments of the detection delay up to order r, when the prior distribution π is heavy-tailed (i.e., when $\mu = 0$) and the r-complete version of the SLLN holds. In the case where $\mu > 0$ (i.e., the prior distribution has an exponential tail) the SR-r procedure is not optimal. This can be expected since it uses the improper uniform prior in the detection statistic.

The same asymptotic results (73)-(74) are true for the CUSUM procedure $\mathsf{T}_{\mathsf{CS}}(a)$ defined in (55) if threshold $a = a_\alpha$ is so selected that $\mathsf{PFA}_\pi(\mathsf{T}_{\mathsf{CS}}(a_\alpha)) \le \alpha$ and $a_\alpha \sim |\log \alpha|$ as $\alpha \to 0$ and the uniform r-complete convergence condition (72) holds.

Hence, the r-complete convergence of the LLR process is the sufficient condition for uniform asymptotic optimality of several popular change detection procedures in class $\mathbb{C}_\pi(\alpha)$.

Complete Convergence and General Non-Bayesian Changepoint Detection Theory

Consider now the non-Bayesian problem assuming that the change point ν is an unknown deterministic number. We focus on the most interesting for applications uniform optimality criterion (67) that requires minimizing the conditional expected delay to detection $\mathsf{CEDD}_\nu(T) = \mathsf{E}_\nu[T - \nu | T > \nu]$ for all values of the change point $\nu \in \mathbb{Z}_+$ in the class of change detection procedures $\mathbb{C}_{\mathrm{PFA}}(m, \beta)$ defined in (66). Recall that this class includes change detection procedures with the maximal local probability of false alarm in the time window m,

$$\mathsf{MLPFA}(T) = \sup_{\ell \ge 0} \mathsf{P}_\infty(T \le \ell + m | T > \ell),$$

which does not exceed the prescribed value $\beta \in (0, 1)$. However, the exact solution to this challenging problem is unknown even in the i.i.d. case.

So instead consider the following asymptotic problem assuming that the given MLPFA β goes to zero: Find a change detection procedure $T^\star$ which minimizes the expected detection delay $\mathsf{E}_\nu[T - \nu | T > \nu]$ asymptotically to first order as $\beta \to 0$. That is, the goal is to design such a detection procedure $T^\star$ that

$$\inf_{T \in \mathbb{C}_{\mathrm{PFA}}(m,\beta)} \mathsf{E}_\nu[T - \nu | T > \nu] = \mathsf{E}_\nu[T^\star - \nu | T^\star > \nu](1 + o(1)) \quad \text{for all } \nu \in \mathbb{Z}_+ \text{ as } \beta \to 0.$$

More generally, we may focus on the asymptotic problem of minimizing moments of the detection delay of order $r \ge 1$:

$$\inf_{T \in \mathbb{C}_{\mathrm{PFA}}(m,\beta)} \mathsf{E}_\nu[(T - \nu)^r | T > \nu] = \mathsf{E}_\nu[(T^\star - \nu)^r | T^\star > \nu](1 + o(1)) \quad \text{for all } \nu \in \mathbb{Z}_+ \text{ as } \beta \to 0.$$

To solve this problem we need to assume that the window length $m = m_\beta$ is a function of the MLPFA constraint β and that m_β goes to infinity as $\beta \to 0$ with a certain appropriate rate. Using [55] the following results can be established.

Let $r \ge 1$ and assume that the complete version of the SLLN holds with some number $0 < I < \infty$, i.e., $n^{-1}\lambda^\nu_{\nu+n}$ converges to I uniformly completely as $n \to \infty$ under P_ν. If $m_\beta = O(|\log \beta|^2)$ as $\beta \to \infty$ and threshold $A = A_\beta$ in the SR procedure is so selected that $\mathsf{MPFA}(\mathsf{T}_{\mathsf{SR}}(A_\beta)) \le \beta$ and $\log A_\beta \sim |\log \beta|$ as $\beta \to 0$, e.g., as defined in [55], then as $\beta \to 0$

$$\inf_{T \in \mathbb{C}_{\mathrm{PFA}}(m_\beta,\beta)} \mathsf{E}_\nu\left[(T - \nu)^r \,|\, T > \nu\right] \sim \left(\frac{|\log \beta|}{I}\right)^r \sim \mathsf{E}_\nu\left[(\mathsf{T}_{\mathsf{SR}} - \nu)^r \,|\, \mathsf{T}_{\mathsf{SR}} > \nu\right] \quad \text{for all } \nu \in \mathbb{Z}_+.$$

A similar result also holds for the CUSUM procedure $\mathsf{T}_{\mathsf{CS}}(a)$ if threshold $a = a_\beta$ is so selected that $\mathsf{MPFA}(\mathsf{T}_{\mathsf{CS}}(a_\beta)) \le \beta$ and $a_\beta \sim |\log \beta|$ as $\beta \to 0$ and the complete version of the SLLN holds for the normalized LLR $n^{-1}\lambda^\nu_{\nu+n}$ as $n \to \infty$.

Hence, the r-complete convergence of the LLR process is the sufficient condition for uniform asymptotic optimality of SR and CUSUM change detection procedures with respect to moments of the detection delay of order r in class $\mathbb{C}_{\mathrm{PFA}}(m_\beta, \beta)$.

4. Quick and Complete Convergence for Markov and Hidden Markov Models

Usually, in particular problems, verification of the SLLN for the LLR process is relatively easy. However, in practice, verifying strengthened r-complete or r-quick versions of the SLLN, i.e., checking condition (72) can cause some difficulty. Many interesting examples where this verification was performed can be found in [5,6]. However, it is interesting to find sufficient conditions for r-complete convergence for a relatively large class of stochastic models.

In this section, we outline this issue for Markov and hidden Markov models based on the results obtained by Pergamenchtchikov and Tartakovsky [55] for ergodic Markov processes and by Fuh and Tartakovsky [56] for hidden Markov models (HMM). See also Tartakovsky [5, Ch 3].

Let $\{X_n\}_{n\in\mathbb{Z}_+}$ be a time-homogeneous Markov process with values in a measurable space $(\mathscr{X},\mathscr{B})$ with the transition probability $P(x,A)$. Let E_x denote the expectation with respect to this probability. Assume that this process is geometrically ergodic, i.e., there exist positives constants $0<R<\infty$, $\kappa>0$, probability measure $\varkappa$ on $(\mathscr{X},\mathscr{B})$ and the Lyapunov $\mathscr{X}\to[1,\infty)$ function V with $\varkappa(V)<\infty$, such that

$$\sup_{n\in\mathbb{Z}_+} e^{\kappa n} \sup_{0<\psi\le V} \sup_{x} \frac{1}{V(x)} \left|\mathsf{E}_x[\psi(X_n)]-\varkappa(\psi)\right| \le R.$$

In the change detection problem, the sequence $\{X_n\}_{n\in\mathbb{Z}_+}$ is a Markov process, such that $\{X_n\}_{1\le n\le\nu}$ is a homogeneous process with the transition density $g(y|x)$ and $\{X_n\}_{n>\nu}$ is homogeneous positive ergodic with the transition density $f_(y|x)$ and the ergodic (stationary) distribution $\varkappa$. In this case, the LLR process λ_n^k can be represented as

$$\lambda_n^k=\sum_{t=k+1}^{n} G(X_t,X_{t-1}),\quad n>k,$$

where $G(y,x)=\log[f(y|x)/g(y|x)]$.

Define

$$I=\int_{\mathscr{X}}\left\{\int_{\mathscr{X}} G(y,x)\,f(y|x)\,\mathrm{d}y\right\}\varkappa(\mathrm{d}x).$$

Under a set of quite sophisticated sufficient conditions the LLR λ_{k+n}^n/n converges r-completely to I (cf. [55]). We omit the details and only mention that the main condition is the finiteness of $(r+1)$-th moment of the LLR increment, $\mathsf{E}_0[(G(X_1,X_0))^{r+1}]<\infty$.

Consider now the HMM with finite state space. Then again, as in the pure Markov case, the main condition for r-complete convergence of λ_{k+n}^n/n to I , where I is specified in Fuh and Tartakovsky [56], is $\mathsf{E}_0[(\lambda_1^0)^{r+1}]<\infty$. Further details can be found in [56].

Similar results for Markov and hidden Markov models hold for the hypothesis testing problem considered in Section 3.1. Specifically, if in the Markov case we assume that the observed Markov process $\{X_n\}_{n\in\mathbb{Z}_+}$ is a time-homogeneous geometrically ergodic with transition density $f_i(y|x)$ under hypothesis H_i $(i=0,1,\dots,N)$ and invariant distribution $\varkappa_i$, then the LLR processes are

$$\lambda_{ij}(n)=\sum_{t=1}^{n} G_{ij}(X_t,X_{t-1}),\quad i,j=0,1,\dots,N,\ i\neq j,$$

where $G_{ij}(y,x)=\log[f_i(y|x)/f_j(y|x)]$. If $\mathsf{E}_i[(G_{ij}(X_1,X_0))^{r+1}]<\infty$ then the LLR $n^{-1}\lambda_{ij}(n)$ converges r-completely to a finite number

$$I_{ij}=\int_{\mathscr{X}}\left\{\int_{\mathscr{X}} G_{ij}(y,x)\,f_i(y|x)\,\mathrm{d}y\right\}\varkappa_i(\mathrm{d}x).$$

doi:10.20944/preprints202305.0904.v1

5. Conclusion

We show that the strengthened versions of the SLLN, specifically the r-quick and r-complete versions, are useful tools for many statistical problems for general non-i.i.d. stochastic models. In particular, r-quick and r-complete convergences for log-likelihood ratio processes are sufficient for near optimality of sequential hypothesis tests and changepoint detection procedures for models with dependent and non-identically distributed observations. Such non-i.i.d. models are typical for modern large-scale information and physical systems that produce Big Data in numerous practical applications. Readers interested in specific applications may find detailed discussions in [4–6,8,19,22,23,34,36,38,54, 55,55–59].

Short Biography of the Author

Alexander G Tartakovsky received the Ph.D. degree in statistics and information theory and the advanced D.Sc. degree from the Moscow Institute of Physics and Technology (PhysTech), Russia, in 1981 and 1990, respectively.

From 1981 to 1992, he was first a Senior Research Scientist and then the Department Head at the Moscow Institute of Radio Technology and a Professor at PhysTech, where he worked on the application of statistical methods to the optimization and modeling of information systems. From 1993 to 1996, he was a Professor at the University of California, Los Angeles (UCLA), first with the Department of Electrical Engineering and then with the Department of Mathematics. From 1997 to 20013, he was a Professor at the Department of Mathematics and an Associate Director of the Center for Applied Mathematical Sciences, University of Southern California (USC). In the late 1990s, he organized one of America's first master's programs in Mathematical Finance (a joint program of the Mathematics and Economics departments at USC). From 2013 to 2015, he was a Professor of statistics with the Department of Statistics at the University of Connecticut, Storrs. From 2016 to 2021, he was the Head of the Space Informatics Laboratory at PhysTech. He is currently the President of AGT StatConsult, Los Angeles, CA, USA. Dr. Tartakovsky also served as visiting faculty at various universities such as Universite de Rouen, France; University of Technology, Sydney, Australia; The Hebrew University of Jerusalem, Israel; University of North Carolina, Chapel Hill; Columbia University; and Stanford University.

Dr. Tartakovsky is an internationally recognized researcher in theoretical and applied statistics, applied probability, sequential analysis, and changepoint detection. He is the author of three books, several book chapters, and over 100 papers across a range of subjects, including theoretical and applied statistics, applied probability, and sequential analysis. His research focuses on a variety of applications including statistical image and signal processing; video surveillance and object detection and tracking; information integration/fusion; intrusion detection and network security; detection and tracking of malicious activity; mathematical/engineering finance applications; pharmacokinetics/ pharmacodynamics; and early detection of epidemics using changepoint methods. Dr. Tartakovsky has provided statistical consulting and developed algorithms and software for many companies and U.S. federal agencies.

Dr. Tartakovsky is a Fellow of the Institute of Mathematical Statistics (IMS) and Senior Member of IEEE. He is an Award-Winning Statistician. He received numerous awards for his work, including the Abraham Wald Prize in Sequential Analysis. He presented several keynote and plenary talks at leading conferences.

References

1. Hsu, P.L.; Robbins, H. Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the United States of America* **1947**, *33*, 25–31.
2. Baum, L.E.; Katz, M. Convergence rates in the law of large numbers. *Transactions of the American Mathematical Society* **1965**, *120*, 108–123.
3. Strassen, V. Almost sure behavior of sums of independent random variables and martingales. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, June 21–July 18, 1965 and December 27, 1965–January 7, 1966*; Le Cam, L.M.; Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; Vol. 2: Contributions to Probability Theory. Part 1, pp. 315–343.
4. Tartakovsky, A.G. Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case. *Statistical Inference for Stochastic Processes* **1998**, *1*, 265–295.
5. Tartakovsky, A.G. *Sequential Change Detection and Hypothesis Testing: General Non-i.i.d. Stochastic Models and Asymptotically Optimal Rules*; Monographs on Statistics and Applied Probability 165, Chapman & Hall/CRC Press, Taylor & Francis Group: Boca Raton, London, New York, 2020.
6. Tartakovsky, A.G.; Nikiforov, I.V.; Basseville, M. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*; Monographs on Statistics and Applied Probability 136, Chapman & Hall/CRC Press, Taylor & Francis Group: Boca Raton, London, New York, 2015.
7. Lai, T.L. On *r*-quick convergence and a conjecture of Strassen. *Annals of Probability* **1976**, *4*, 612–627.
8. Lai, T.L. Asymptotic optimality of invariant sequential probability ratio tests. *Annals of Statistics* **1981**, *9*, 318–333.
9. Chow, Y.S.; Lai, T.L. Some one-sided theorems on the tail distribution of sample sums with applications to the last time and largest excess of boundary crossings. *Transactions of the American Mathematical Society* **1975**, *208*, 51–72.
10. Fuh, C.D.; Zhang, C.H. Poisson equation, moment inequalities and quick convergence for Markov random walks. *Stochastic Processes and Their Applications* **2000**, *87*, 53–67.
11. Wald, A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **1945**, *16*, 117–186.
12. Wald, A. *Sequential Analysis*; John Wiley & Sons, Inc: New York, USA, 1947.
13. Wald, A.; Wolfowitz, J. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* **1948**, *19*, 326–339.
14. Burkholder, D.L.; Wijsman, R.A. Optimum properties and admissibility of sequential tests. *Annals of Mathematical Statistics* **1963**, *34*, 1–17.
15. Matthes, T.K. On the optimality of sequential probability ratio tests. *Annals of Mathematical Statistics* **1963**, *34*, 18–21.
16. Ferguson, T.S. *Mathematical Statistics: A Decision Theoretic Approach*; Probability and Mathematical Statistics, Academic Press, 1967.
17. Lehmann, E.L. *Testing Statistical Hypotheses*; John Wiley & Sons, Inc: New York, USA, 1968.
18. Shiryaev, A.N. *Optimal Stopping Rules*; Vol. 8, *Series on Stochastic Modelling and Applied Probability*, Springer-Verlag: New York, USA, 1978.
19. Tartakovsky, A.G. *Sequential Methods in the Theory of Information Systems*; Radio i Svyaz': Moscow, RU, 1991. In Russian.
20. Golubev, G.K.; Khas'minskii, R.Z. Sequential testing for several signals in Gaussian white noise. *Theory of Probability and its Applications* **1984**, *28*, 573–584.
21. Tartakovsky, A.G. Asymptotically optimal sequential tests for nonhomogeneous processes. *Sequential Analysis* **1998**, *17*, 33–62.
22. Verdenskaya, N.V.; Tartakovskii, A.G. Asymptotically optimal sequential testing of multiple hypotheses for nonhomogeneous Gaussian processes in an asymmetric situation. *Theory of Probability and its Applications* **1991**, *36*, 536–547.
23. Fellouris, G.; Tartakovsky, A.G. Multichannel sequential detection – Part I: Non-i.i.d. data. *IEEE Transactions on Information Theory* **2017**, *63*, 4551–4571. https://doi.org/10.1109/TIT.2017.2689785.

24. Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society - Series B Methodology* **1950**, *12*, 137–144.
25. Chernoff, H. Sequential design of experiments. *Annals of Mathematical Statistics* **1959**, *30*, 755–770.
26. Kiefer, J.; Sacks, J. Asymptotically optimal sequential inference and design. *Annals of Mathematical Statistics* **1963**, *34*, 705–750.
27. Lorden, G. Integrated risk of asymptotically Bayes sequential tests. *Annals of Mathematical Statistics* **1967**, *38*, 1399–1422.
28. Lorden, G. Nearly-optimal sequential tests for finitely many parameter values. *Annals of Statistics* **1977**, *5*, 1–21.
29. Pavlov, I.V. Sequential procedure of testing composite hypotheses with applications to the Kiefer-Weiss problem. *Theory of Probability and its Applications* **1990**, *35*, 280–292.
30. Baron, M.; Tartakovsky, A.G. Asymptotic optimality of change-point detection schemes in general continuous-time models. *Sequential Analysis* **2006**, *25*, 257–296. Invited Paper in Memory of Milton Sobel.
31. Mosteller, F. A *k*-sample slippage test for an extreme population. *Annals of Mathematical Statistics* **1948**, *19*, 58–65.
32. Bakut, P.A.; Bolshakov, I.A.; Gerasimov, B.M.; Kuriksha, A.A.; Repin, V.G.; Tartakovsky, G.P.; Shirokov, V.V. *Statistical Radar Theory*; Vol. 1 (G. P. Tartakovsky, Editor), Sovetskoe Radio: Moscow, USSR, 1963. In Russian.
33. Basseville, M.; Nikiforov, I.V. *Detection of Abrupt Changes – Theory and Application*; Information and System Sciences Series, Prentice-Hall, Inc: Englewood Cliffs, NJ, USA, 1993. Online.
34. Jeske, D.R.; Steven, N.T.; Tartakovsky, A.G.; Wilson, J.D. Statistical methods for network surveillance. *Applied Stochastic Models in Business and Industry* **2018**, *34*, 425–445. Discussion Paper, https://doi.org/10.1002/asmb.2326.
35. Jeske, D.R.; Steven, N.T.; Wilson, J.D.; Tartakovsky, A.G. Statistical network surveillance. *Wiley StatsRef: Statistics Reference Online* **2018**, pp. 1–12. https://doi.org/https://doi.org/10.1002/9781118445112.stat08055.
36. Tartakovsky, A.G.; Brown, J. Adaptive spatial-temporal filtering methods for clutter removal and target tracking. *IEEE Transactions on Aerospace and Electronic Systems* **2008**, *44*, 1522–1537.
37. Szor, P. *The Art of Computer Virus Research and Defense*; Addison-Wesley Professional: Upper Saddle River, NJ, USA, 2005.
38. Tartakovsky, A.G. Rapid detection of attacks in computer networks by quickest changepoint detection methods. In *Data Analysis for Network Cyber-Security*; Adams, N.; Heard, N., Eds.; Imperial College Press: London, UK, 2014; pp. 33–70.
39. Tartakovsky, A.G.; Rozovskii, B.L.; Blaźek, R.B.; Kim, H. Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology* **2006**, *3*, 252–293.
40. Tartakovsky, A.G.; Rozovskii, B.L.; Blaźek, R.B.; Kim, H. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing* **2006**, *54*, 3372–3382.
41. Siegmund, D. Change-points: from sequential detection to biology and back. *Sequential Analysis* **2013**, *32*, 2–14.
42. Moustakides, G.V. Sequential change detection revisited. *Annals of Statistics* **2008**, *36*, 787–807.
43. Page, E.S. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–114.
44. Shiryaev, A.N. On optimum methods in quickest detection problems. *Theory of Probability and its Applications* **1963**, *8*, 22–46.
45. Moustakides, G.V.; Polunchenko, A.S.; Tartakovsky, A.G. A numerical approach to performance analysis of quickest change-point detection procedures. *Statistica Sinica* **2011**, *21*, 571–596.
46. Moustakides, G.V.; Polunchenko, A.S.; Tartakovsky, A.G. Numerical comparison of CUSUM and Shiryaev–Roberts procedures for detecting changes in distributions. *Communications in Statistics – Theory and Methods* **2009**, *38*, 3225–3239.
47. Lorden, G. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics* **1971**, *42*, 1897–1908.
48. Moustakides, G.V. Optimal stopping times for detecting changes in distributions. *Annals of Statistics* **1986**, *14*, 1379–1387.
49. Pollak, M. Optimal detection of a change in distribution. *Annals of Statistics* **1985**, *13*, 206–227.

50. Tartakovsky, A.G.; Pollak, M.; Polunchenko, A.S. Third-order asymptotic optimality of the generalized Shiryaev–Roberts changepoint detection procedures. *Theory of Probability and its Applications* **2012**, *56*, 457–484. https://doi.org/10.1137/S0040585X97985534.
51. Polunchenko, A.S.; Tartakovsky, A.G. On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *Annals of Statistics* **2010**, *38*, 3445–3457.
52. Shiryaev, A.N. The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathematics – Doklady* **1961**, *2*, 795–799. Translation from Doklady Akademii Nauk SSSR, **138**:1039–1042, 1961.
53. Tartakovsky, A.G. Discussion on "Is Average Run Length to False Alarm Always an Informative Criterion?" by Yajun Mei. *Sequential Analysis* **2008**, *27*, 396–405.
54. Liang, Y.; Tartakovsky, A.G.; Veeravalli, V.V. Quickest change detection with non-stationary post-change observations. *IEEE Transactions on Information Theory* **2023**, *69*, 3400–3414. https://doi.org/0.1109/TIT.2022.3230583.
55. Pergamenchtchikov, S.; Tartakovsky, A.G. Asymptotically optimal pointwise and minimax quickest change-point detection for dependent data. *Statistical Inference for Stochastic Processes* **2018**, *21*, 217–259.
56. Fuh, C.D.; Tartakovsky, A.G. Asymptotic Bayesian theory of quickest change detection for hidden Markov models. *IEEE Transactions on Information Theory* **2019**, *65*, 511–529. https://doi.org/10.1109/TIT.2018.2843379.
57. Kolessa, A.; Tartakovsky, A.; Ivanov, A.; Radchenko, V. Nonlinear estimation and decision-making methods in short track identification and orbit determination problem. *IEEE Transactions on Aerospace and Electronic Systems* **2020**, *56*, 301–312. https://doi.org/10.1109/TAES.2019.2911760.
58. Tartakovsky, A.; Berenkov, N.; Kolessa, A.; Nikiforov, I. Optimal sequential detection of signals with unknown appearance and disappearance points in time. *IEEE Transactions on Signal Processing* **2021**, *69*, 2653–2662. https://doi.org/10.1109/TSP.2021.3071016.
59. Pergamenchtchikov, S.M.; Tartakovsky, A.G.; Spivak, V.S. Minimax and pointwise sequential changepoint detection and identification for general stochastic models. *Journal of Multivariate Analysis* **2022**, *190*, 1–22.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.