2024 60th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3315-4103-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ALLERTON63246.2024.10735265

Quickest Change Detection Using Mismatched CUSUM

Extended Abstract

Austin Cooper and Sean Meyn*

I. Introduction

The field of quickest change detection (QCD) concerns design and analysis of algorithms to estimate the time at which an important event takes place and identify properties of the post-change behavior.

The goal of the researched surveyed here is to devise a stopping time adapted to the observations that minimizes an L_1 loss. Approximately optimal solutions are well known under a variety of assumptions. In the work surveyed here we consider the CUSUM statistic, which evolves as a one-dimensional reflected random walk driven by a functional of the observations. It is known that the optimal functional is a log likelihood ratio subject to special statistical assumptions.

Among the questions considered in current research are, 1. What is the performance for a given functional of the observations in Bayesian and minimax settings? 2. How do the conclusions change when there is dependency between pre- and post-change behavior? 3. How can techniques from statistics and machine learning be adapted to approximate the best functional in a given class?

This survey focuses on topics 1 and 3, whereas topic 2 is addressed in [2]. Analysis is cast entirely within the Bayesian setting.

QCD model The general QCD model includes a sequence of observations $\boldsymbol{Y} := \{Y_k : k \geq 0\}$ evolving in an abstract set Y (typically a subset of Euclidean space). The onset of anomalous behavior (the "change time") is denoted τ_{a}; this is formalized through the representation

$$Y_k = X_k^0 \mathbf{1}_{k<\tau_{\mathsf{a}}} + X_k^1 \mathbf{1}_{k\geq\tau_{\mathsf{a}}}\,, \qquad k \geq 0\,. \tag{1}$$

In general this is simply a notational convention: $X_k^0 := Y_k$ for $k < \tau_{\mathsf{a}}$, and $X_k^1 := Y_k$ otherwise. An estimate of the change time, denoted τ_{s}, is assumed adapted to the observations: $U_k := \mathbf{1}\{\tau_{\mathsf{s}} = k\} = \phi_k(Y_0, \ldots, Y_k)$, $k \geq 0$, for a family of functions $\{\phi_k\}$.

The vast majority of theory requires statistical independence of $\boldsymbol{X}^0$, $\boldsymbol{X}^1$ and τ_{a}. This is the case in Shiryaev's conditional i.i.d. model, for which the stochastic processes $\boldsymbol{X}^0$, $\boldsymbol{X}^1$ are also assumed i.i.d.; an optimal policy is easily described if in addition the distribution of the change is geometric [11].

We choose a criterion for optimality that reflects our desire to trigger an alarm briefly before the change time: for given $\kappa > 0$, and a policy ϕ, denote

$$J(\phi) = \mathsf{E}\big[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_+ + \kappa(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-\big] \tag{2}$$

with $x_+ = \max(x, 0)$, $x_- = \max(-x, 0)$. Motivation for inclusion of the *cost of eagerness* $\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-]$ is clear when $\boldsymbol{X}^0$ and τ_{a} are statistically dependent [2].

Most successful approaches to QCD begin with the construction of a real-valued stochastic process $\{\mathcal{X}_n\}$, adapted to the observations, and the stopping rule is of the threshold form,

$$\tau_{\mathsf{s}} = \min\{n \geq 0 : \mathcal{X}_n \geq \mathrm{H}\}\,, \tag{3}$$

with $\mathrm{H} > 0$. Two famous examples are found in the test of Shiryaev–Roberts, and Page's CUSUM. The latter is the focus of this paper, in which $\{\mathcal{X}_n\}$ is defined as a reflected random walk,

$$\mathcal{X}_{n+1} = \max\{0, \mathcal{X}_n + F_{n+1}\} \tag{4}$$

initialized with $\mathcal{X}_0 = 0$, where $\{F_{n+1} : n \geq 0\}$ is a stochastic process adapted to the observations.

Contributions Performance of the CUSUM test is approximated in the asymptotic setting in which $\kappa \uparrow \infty$ (hence a strong penalty for false alarm). Analytical techniques rooted in large deviations theory lead to approximations of the optimal threshold and cost as a function of κ, even in non-ideal settings. These approximations lend themselves to optimization of the statistic F_n over a finite-dimensional function class.

This general theory also motivates application of reinforcement learning techniques to estimate a near-optimal policy. One theme of [4], [3] is the application of observation-driven statistics such as (4) to form a "surrogate" information state for this purpose.

Literature See [9], [10] for excellent recent surveys on QCD theory. Much of past research is restricted to the conditionally i.i.d. model. Recent extensions to conditionally Markov models or hidden Markov models is found in [15], [14].

*AC and SM are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA; austin.cooper,meyn @ufl.edu

Financial support from NSF awards CCF-2306023 and DMS-2427265 is gratefully acknowledged.

II. Assumptions and Main Results

In the main results surveyed here we adopt the observation model (1) in which $\boldsymbol{X}^0$, $\boldsymbol{X}^1$ are mutually independent stationary stochastic processes, and independent of the change time τ_{a}. The marginal distributions are denoted π^0, π^1 respectively.

We assume that Y is equipped with a sigma-algebra $\mathcal{B}(\mathsf{Y})$; typically $\mathcal{B}(\mathsf{Y})$ is the Borel sigma-algebra when Y is a topological space. With $\mathbb{R}^m$ equipped with the Borel sigma-algebra, if $G\colon \mathsf{Y} \to \mathbb{R}^m$ is a measurable vector-valued function we denote $\pi^i(G) = \int G(y)\,\pi^i(dy)$ for $i = 0, 1$.

Our interest is approximating the performance of the CUSUM test, and also approximating the optimal threshold for a given value of κ. The following two assumptions are in place throughout:

A1. We consider $F_n = F(Y_n)$ for a measurable function $F\colon \mathsf{Y} \to \mathbb{R}$. Letting $m_i = \pi_i(F)$ for $i = 0, 1$, it is assumed that $m_0 < 0$ and $m_1 > 0$.

A2. *Regular geometric tail:* for some $\varrho_{\mathsf{a}} \in (0, \infty)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}\{\tau_{\mathsf{a}} \ge n\} = -\varrho_{\mathsf{a}} \tag{5}$$

Under (A1), it follows that $\{\mathcal{X}_n\}$ evolves as a reflected random walk (RRW) with negative drift for $n < \tau_{\mathsf{a}}$ and thereafter a RRW with positive drift.

A. Examples

The numerical results obtained in our recent work are restricted to three settings:

Conditional i.i.d. model This is the classical setting in which $\boldsymbol{X}^0$, $\boldsymbol{X}^1$ are i.i.d.. Consequently, $\{\mathcal{X}_n\}$ evolves as the workload in a GI/G/1 queue, with a change in load parameter at the change time τ_{a} (e.g. [1], [5]).

Suppose that the marginals are mutually absolutely continuous with log likelihood ratio (LLR) denoted $L = \log(d\pi^1/d\pi^0)$. It is known that the use of $F_n = L(Y_n)$ in (4) defines a test that is approximately optimal under certain performance criteria [13]. This function satisfies the desired sign conventions:

$$\pi^0(L) = -D(\pi^0\|\pi^1) < 0\,, \quad \pi^1(L) = D(\pi^1\|\pi^0) > 0$$

where D denotes relative entropy.

Markov model When there is memory in the observations it is necessary to perform some transformation to justify the use of $F_n = F(Y_n)$. In the Markovian setting, for each $i = 0, 1$ we take $X_k^i = (\Phi_{k-1}^i; \Phi_k^i)$ in which $\boldsymbol{\Phi}^i$ is a stationary Markov chain; its transition kernel is denoted P_i, and invariant measure ϖ_i, so that $\Phi_k^i \sim \varpi_i$ for all k. It follows that $X_k^i \sim \pi^i$ for each i, k, with $\pi^i(dx, dz) = \varpi_i(dx)P_i(x, dz)$.

Suppose that there are transition densities $\{g_0, g_1\}$ with respect to some reference measure μ:

$$P_0(x, dz) = g_0(x, z)\mu(dz)\,, \;\; P_1(x, dz) = g_1(x, z)\mu(dz)$$

Then, the LLR of the transition densities is denoted

$$L_\infty(x, z) = \log\Big(\frac{g_1(x, z)}{g_0(x, z)}\Big) \tag{6}$$

Once again the function $F^* = L_\infty$ is approximately optimal in certain settings [13]. Moreover, the sign conventions in (A1) hold:

$$\pi^0(F^*) = -K(P_0\|P_1) < 0\,, \;\; \pi^1(F^*) = K(P_1\|P_0) > 0$$

where K denotes the Donsker-Varadhan rate function. In particular,

$$K(P_1\|P_0) = \int L_\infty(x, z)\,\varpi_1(dx)P_1(x, dz) \tag{7}$$

Hidden Markov Model Suppose that the observations are a function of a Markov chain prior to the change time, $Y_k = h(\Phi_k^0)$ for $k < \tau_{\mathsf{a}}$. To apply techniques from the fully observed Markov setting, the observations might be replaced by the conditional distributions as in treatment of partially observed Markov Decision Processes (MDP). This leads to a highly complex test that cannot be approximated without a model of the observations. Instead, in Section III we show how to obtain the best function F^* within a finite dimensional function class.

B. Assumptions on CUSUM statistic

It is assumed throughout that (A1) and (A2) hold, so in particular $F_n = F(Y_n)$ for each n. In this subsection we summarize the remaining assumptions required in analysis. In particular, many of the definitions and assumptions that follow are required in application of Large Deviations Theory for RRWs in approximating the cost (2).

If $\boldsymbol{X}^0$, $\boldsymbol{X}^1$ are each i.i.d., then sufficient conditions for the assumptions that follow are easily formulated. If these stochastic processes are Markov chains, then Lyapunov function criteria justifying the assumptions are contained in [7], [8].

Let $\mathcal{G}$ denote a family of real-valued measurable functions for which the *cumulative generating function* (CGF) exists and is finite: for $F \in \mathcal{G}$ and $i = 0, 1$,

$$\Lambda_i(F) = \lim_{n\to\infty} \frac{1}{n} \log \mathsf{E}\Big[\exp\Big(\sum_{k=0}^{n-1} F(X_k^i)\Big)\Big] \tag{8}$$

The additional assumptions summarized in the following are largely restricted to $\boldsymbol{X}^0$.

For $F \in \mathcal{G}$ denote by $\breve{\mu}^0_{(n)}$ the probability measure on $\mathcal{B}(\mathsf{Y}^{n+1})$ satisfying, for $Z \in \mathcal{B}(\mathsf{Y}^{n+1})$,

$$\begin{aligned}\breve{\mu}^0_{(n)}\{Z\} &= \beta_n^{-1}\mathsf{E}\big[\exp(S_n)\mathbf{1}_Z(X_0^0,\dots,X_n^0)\big]\\ S_n &= \sum_{k=0}^{n-1} F(X_k^0)\,,\quad \beta_n = \mathsf{E}[\exp(S_n)]\end{aligned} \tag{9}$$

Its last marginal is denoted $\breve{\pi}^0_{(n)}$:

$$\breve{\pi}^0_{(n)}\{A\} = \beta_n^{-1}\mathsf{E}\big[\exp(S_n)\mathbf{1}_A(X_n^0)\big]\,,\quad A \in \mathcal{B}(\mathsf{Y})$$

For the special case $F \equiv 0$ these probability measures are denoted $\mu^0_{(n)}$, $\pi^0_{(n)}$ respectively; stationarity implies that $\pi^0_{(n)} = \pi^0$ for each n.

The following limits are assumed to exist for $F \in \mathcal{G}$:

Twisted marginal:

$$\breve{\pi}^0\{A\} = \lim_{n\to\infty} \breve{\pi}^0_{(n)}\{A\}\,,\quad A \in \mathcal{B}(\mathsf{Y}) \tag{10}$$

Twisted process: the stationary stochastic process whose finite dimensional marginals are defined for $Z \in \mathcal{B}(\mathsf{Y}^{m+1})$ by

$$\breve{\mu}^0\{Z\} := \lim_{n\to\infty} \breve{\mu}^0_{(n)}\{\mathsf{Y}^{n-m}\times Z\} \tag{11}$$

Relative entropy rate: For $F \in \mathcal{G}$ it is assumed that the following limit exists (though possibly infinite):

$$\mathcal{K}(\breve{\mu}^0\|\mu^0) = \lim_{n\to\infty} \frac{1}{n} D(\breve{\mu}^0_{(n)}\|\mu^0_{(n)}) \tag{12}$$

where D denotes relative entropy. It is also assumed that $\mathcal{K}(\mu^1\|\mu^0) < \infty$ where

$$\mathcal{K}(\mu^1\|\mu^0) = \lim_{n\to\infty} \frac{1}{n} D(\mu^1_{(n)}\|\mu^0_{(n)}) \tag{13}$$

For the Markov model with $F = L_\infty$ we have $\Lambda_0(F) = 0$ and $\mathcal{K}(\mu^1\|\mu^0) = K(P_1\|P_0)$ (see (6) and surrounding discussion). This follows from the more general statement:

Lemma 2.1: $\mathcal{K}(\breve{\mu}^0\|\mu^0) = \breve{\pi}^0(F) - \Lambda_0(F)$. ∎

C. Performance approximations

The assumptions on $F \in \mathcal{G}$ are imposed since we apply theory of rare events for RRWs to approximate the behavior of $\{\mathcal{X}_n : 0 \le n < \tau_{\mathsf{a}}\}$ (e.g., [5]).

We perform a temporal and spatial scaling that is standard in this literature: for a given threshold $\mathrm{H} > 0$ denote by $\{x_t^{(\mathrm{H})} : t \ge 0\}$ the continuous function defined by $x_t^{(\mathrm{H})} = \mathrm{H}^{-1}\mathcal{X}_k$ for $t = k/\mathrm{H}$, and by piecewise linear interpolation for all other $t \ge 0$. When τ_{s} is defined using threshold H, then $\tau_{\mathsf{s}} \le T\mathrm{H}$ if and only if $x_t^{(\mathrm{H})} \ge 1$ for some $t \le T$.

We require a few additional assumptions beyond those summarized in Section II-B. Let $\Upsilon_0\colon \mathbb{R} \to \mathbb{R}\cup\{\infty\}$ denote the convex function defined by $\Upsilon_0(\upsilon) = \Lambda_0(\upsilon F)$ for $\upsilon \in \mathbb{R}$.

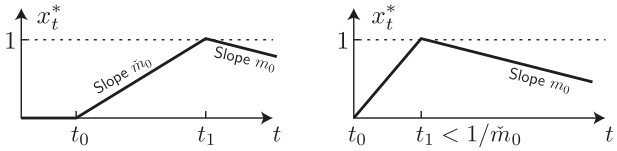

Fig. 1. *Two paths*: $\{x_t^{(\mathrm{H})} : t \ge 0\}$ approximating the path shown on the left is far more likely than the one shown on the right.

A3. *Assumptions imposed for approximations.* $F \in \mathcal{G}$ and the following hold:

- There are solutions $\upsilon_+ > \upsilon_0 > 0$ to

$$\Upsilon_0(\upsilon_0) = 0\,,\quad \Upsilon_0(\upsilon_+) = \varrho_{\mathsf{a}} \tag{14}$$

- The function Υ_0 is finite valued and continuously differentiable in a neighborhood of $[0, \upsilon_+]$.
- $\upsilon F \in \mathcal{G}$ for each υ in a neighborhood of $[0, \upsilon_+]$. In particular, the following limit exists for each $A \in \mathcal{B}(\mathsf{Y})$

$$\breve{\pi}^0_\upsilon(A) = \lim_{n\to\infty} \frac{\mathsf{E}\big[\exp\big(\upsilon S_n\big)\mathbf{1}\{X_n^0 \in A\}\big]}{\mathsf{E}\big[\exp\big(\upsilon S_n\big)\big]} \tag{15}$$

with $S_n = \sum_{k=0}^{n-1} F(X_k^0)$ (recall (10)).

Denote

$$\breve{m}_0 = \Upsilon'(\upsilon_0)\,,\quad \breve{m}_+ = \Upsilon'(\upsilon_+) \tag{16}$$

Under (A3) we have $\Upsilon'(0) = m_0 < 0$ and $\breve{m}_+ > \breve{m}_0 > 0$. The following result follows from sample path large deviations theory from [5].

Lemma 2.2: Suppose that (A1) and (A3) hold, and that $\tau_{\mathsf{a}} = \infty$ with probability one. Then, for $T > 0$,

$$\lim_{\mathrm{H}\to\infty} \frac{1}{\mathrm{H}} \log \mathsf{P}\Big\{\sup_{0\le t\le T} x_t^{(\mathrm{H})} \ge 1\Big\} = -e_0(T)$$

$$\textit{with}\quad e_0(T) = \begin{cases} TI_0(1/T) & T < 1/\breve{m}_0\\ I_0(\breve{m}_0)/\breve{m}_0 & T \ge 1/\breve{m}_0\end{cases}$$

where I_0 is the convex dual of Υ_0.

Fig. 1 illustrates two piecewise linear paths that might approximate the RRW for a very large threshold. The path on the left is more likely because the exponent $e_0(T)$ is minimized when $T \ge 1/\breve{m}_0$. This reflects the well known approximation: the most likely path for a random walk to hit a high level is linear with slope $\breve{m}_0$

Denote by $\bar{J}(\mathrm{H}, \kappa)$ the value of the expectation (2) using CUSUM with threshold $\mathrm{H} > 0$, and

$$\bar{J}^*(\kappa) = \min_{\mathrm{H}} \bar{J}(\mathrm{H}, \kappa)\,,\ \bar{\mathrm{H}}^*(\kappa) = \arg\min_{\mathrm{H}} \bar{J}(\mathrm{H}, \kappa) \tag{17}$$

Approximations for each grow logarithmically in κ:

$$\bar{\mathrm{H}}^*_\infty(\kappa) = \frac{1}{\upsilon_+}\log(\kappa) \tag{18a}$$

$$\bar{J}^*_\infty(\kappa) = \frac{1}{m_1}\frac{1}{\upsilon_+}\log(\kappa) \tag{18b}$$

These approximations are justified in the following. An outline of the proof of Prop. 2.3 may be found in the Appendix.

Proposition 2.3: Suppose (A1)–(A3) hold. Then,

$$\bar{\mathrm{H}}^*(\kappa) = \bar{\mathrm{H}}^*_\infty(\kappa) + o(\log(\kappa))$$
$$\bar{J}^*(\kappa) = \bar{J}^*_\infty(\kappa) + o(\log(\kappa))$$

III. Optimizing CUSUM

We turn next to methods to obtain the best function within a finite dimensional function class of the form $\{F_\theta : \theta \in \Theta\} \subset \mathcal{G}$ with $\Theta \subset \mathbb{R}^d$.

A. In search of stationary points

When the function F_θ is used in CUSUM, the corresponding approximation of the cost is denoted

$$\bar{J}^*_\infty(\kappa;\theta) = \frac{1}{m_1^\theta}\frac{1}{\upsilon_+^\theta}\log\kappa \tag{19}$$

Here we characterize stationary points of this approximation. In Section III-B we find that the conclusions are far more elegant when the function class is linear.

Denote $\Upsilon_0^\theta(\upsilon) = \Lambda_0(\upsilon F_\theta)$ for each $\theta \in \Theta$ and $\upsilon \in \mathbb{R}$. If F_θ satisfies (A3) then we let $\upsilon_+^\theta > \upsilon_0^\theta > 0$ denote the solutions to

$$\Upsilon_0^\theta(\upsilon_0^\theta) = 0\,, \quad \Upsilon_0^\theta(\upsilon_+^\theta) = \varrho_{\mathsf{a}} \tag{20}$$

A4. *Assumptions imposed on function class*. The set Θ is open, with F_θ satisfying (A3) for each $\theta \in \Theta$. Moreover,

(i) $\psi_\theta(y) = \nabla_\theta F_\theta(y)$ exists for each y, θ to define a d-dimensional function whose elements satisfy the assumptions of Section II-B.

(ii) The function $\Upsilon_0^\theta(\upsilon)$ is continuously differentiable for υ, θ in a neighborhood of $\{(\upsilon,\theta) : \upsilon \in [0, \upsilon_+^\theta]\,,\ \theta \in \Theta\}$.

(iii) There is a vector $v \in \mathbb{R}^d$ such that $v^{\intercal}\psi_\theta \equiv 1$.

Under (A4) we let $\breve{\pi}^0_{\upsilon,\theta}$ denote the probability measure (15) obtained using υ with function F_θ. When using υ_+^θ we simplify the notation to $\breve{\pi}^0_{+,\theta}$, and write $\breve{m}_0^{+,\theta} = \breve{\pi}^0_{+,\theta}(F_\theta)$.

Proposition 3.1: Suppose that (A2) and (A4) hold, and that $\theta^\bullet \in \Theta$ is a stationary point: $\nabla_\theta \bar{J}^*_\infty(\kappa;\theta^\bullet) = 0$ for some (and hence all) $\kappa > 0$. Then,

$$\breve{m}_0^{+,\theta^\bullet} = m_1^{\theta^\bullet} \quad \textit{and} \quad \breve{\pi}^0_{+,\theta^\bullet}\big(\psi_{\theta^\bullet}\big) = \pi^1\big(\psi_{\theta^\bullet}\big) \tag{21}$$

Conversely, if (21) holds then $\theta^\bullet$ is a stationary point.

The proof is contained in [2]. The first step is this key identity:

Lemma 3.2: Subject to (A2) and (A4),

$$\begin{aligned}\nabla_\theta \log \bar{J}^*_\infty(\kappa;\theta) &= -\nabla_\theta \log m_1^\theta - \nabla_\theta \log \upsilon_+^\theta \\ &= -\frac{1}{m_1}\pi^1\big(\psi_\theta\big) + \frac{1}{\breve{m}_0^{+,\theta}}\breve{\pi}^0_{+,\theta}\big(\psi_\theta\big)\end{aligned} \tag{22}$$

B. Linear family

If the function class is linear then $F_\theta = \theta^{\intercal}\psi$ so that $\psi_\theta = \psi$, with $\psi_i \in \mathcal{G}$ for each i.

In Prop. 3.3 we show that the minimization over θ of $\bar{J}^*_\infty$ may be cast as a convex program with objective $\Gamma(\theta) = \Gamma_0(\theta) + \frac{1}{2}(v^{\intercal}\theta)^2$, where

$$\Gamma_0(\theta) = \Lambda_0(F_\theta) - \pi^1(F_\theta) \tag{23}$$

This objective function Γ is strictly convex when the $d \times d$ autocorrelation matrix R_θ is full rank, with entries,

$$R_\theta(i,j) := \breve{\pi}^0_{1,\theta}(\psi^i\psi^j) \tag{24}$$

Proposition 3.3: Suppose that (A2) and (A4) hold. Let $\theta^* = \theta^\circ + r^\circ v$, where θ° minimizes Γ and $r^\circ = \varrho_{\mathsf{a}} - \Lambda_0(F_{\theta^\circ})$. Then, θ^* is a global minimizer of $\bar{J}^*_\infty$.

Outline of proof: The first step in the proof is to recognize that Γ_0 is not strictly convex, because for any $\theta \in \mathbb{R}^d$ and $r \in \mathbb{R}$ we have

$$\Gamma_0(\theta + rv) = \Lambda_0(F_\theta + r) - \pi^1(F_\theta + r) = \Gamma_0(\theta)$$

In particular, $v^{\intercal}\nabla\Gamma_0(\theta) = 0$ for any θ, and $v^{\intercal}\theta^\circ = 0$.

It follows that $0 = \nabla\Gamma(\theta^\circ) = \nabla\Gamma_0(\theta^\circ)$, which is equivalently expressed $\breve{\pi}^0_{1,\theta^\circ}(\psi_{\theta^\circ}) = \pi^1(\psi_{\theta^\circ})$; the identity $\breve{m}_0^{1,\theta^\circ} = m_1^{\theta^\circ}$ follows since the function class is linear. These conclusions imply (21) only if $\upsilon_+^{\theta^\circ} = 1$.

The vector θ^* satisfies the same two identities, and $\upsilon_+^{\theta^*} = 1$ by construction:

$$\Lambda_0(F_{\theta^*}) = \Lambda_0(F_{\theta^\circ + r^\circ v}) = \Lambda_0(F_{\theta^\circ}) + r^\circ = \varrho_{\mathsf{a}}$$

Consequently (21) holds, so that θ^* is a stationary point of $\bar{J}^*_\infty$ due to Prop. 3.1.

More work is required to establish optimality. ∎

Information theoretic representations For the linear function class we have seen that we can take without loss of generality $\upsilon_+^\theta = 1$, and then (19) simplifies to $\bar{J}^*_\infty(\kappa;\theta) = \log(\kappa)/m_1^\theta$. The definition $m_1^\theta := \pi^1(F_\theta)$ and calculations similar to those used to establish Lemma 2.1 lead to the following:

Proposition 3.4: Under (A4),
$m_1^\theta = \Lambda_0(F_\theta) + \mathcal{K}(\mu^1\|\mu^0) - \mathcal{K}(\mu^1\|\breve{\mu}^0_\theta)$.

Consequently, maximizing m_1^θ subject to $\Lambda_0(F_\theta) = \varrho_{\mathsf{a}}$ is equivalent to minimizing $\mathcal{K}(\mu^1\|\breve{\mu}^0_\theta)$ subject to the same constraint.

Example: Optimization for the Markov model. Recall the definition of L_∞ in (6). We obtain $\Lambda_0(F_{\theta^*}) = \varrho_{\mathsf{a}}$ and $\mathcal{K}(\mu^1\|\breve{\mu}^0_{\theta^*}) = 0$ using $F_{\theta^*} = L_\infty + \varrho_{\mathsf{a}}$ (provided this is in the function class).

Example: scalar offset. Consider optimization of a scalar offset, so that $F_\theta = F + \theta$ for $\theta \in \mathbb{R}$. In this case $\psi_\theta(y) = 1$ for all y, so that we can take $v = 1$ in (A4). Applying Prop. 3.1 we conclude that the optimal parameter is characterized by a single mean constraint:

Proposition 3.5: Assume that the function $F + r$ satisfies (A1) and (A3) for some r. Then an optimal scalar offset is a solution to $m_1^{\theta^*} = \breve{m}_0^{+,\theta^*}$. ∎

The proposition leaves out methods to compute θ^*. We have $m_1^\theta = m_1 + \theta$, but dependency of $\breve{m}_0^{+,\theta}$ on θ is far more complex. This is resolved by enlarging the function class so that Prop. 3.3 is applicable.

Proposition 3.6: Consider the basis $\psi = (F; 1)$, so that $F_\theta = \theta_1 F + \theta_2$, and assume that F satisfies the assumptions of Prop. 3.5. Then, an optimizer of $\bar{J}_\infty^*$ is of the form $\theta^* = \theta^\circ + r^\circ(0; 1)$, obtained as follows:

1. $\theta_1^\circ = \arg\min_\upsilon [\Lambda_0(\upsilon F) - \upsilon\pi^1(F)]$ and $\theta_2^\circ = 0$.
2. $r^\circ = \varrho_{\mathsf{a}} - \Lambda_0(F_{\theta^\circ})$. ∎

The optimizer is not unique: if the function F^* is optimal within a linear function class, then so is kF^* for any $k > 0$. Here are two explanations:

- We can scale the threshold by the same constant to obtain an equivalent test.
- The approximation (19) is constant under scalings: $m_1^{\theta/k}\upsilon_+^{\theta/k} = m_1^\theta \upsilon_+^\theta$ for all $k > 0$.

In the setting of Prop. 3.6, on choosing $k = 1/\theta_1^\circ > 0$ the function below is optimal for the problem considered in Prop. 3.5:

$$F^* = F + [\varrho_{\mathsf{a}} - \Lambda_0(\theta_1^\circ F)]/\theta_1^\circ \tag{25}$$

with $\upsilon_+ = \theta_1^\circ$ rather than unity.

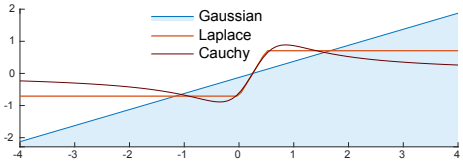

Fig. 2. Three choices for F

C. Numerical Experiments

See [4], [3] for results for the conditionally i.i.d. model. One example is described briefly here, with the following specifications:

- $\boldsymbol{X}^i$ i.i.d. Gaussian on $\mathbb{R}$, with a change in mean at time τ_{a}. Hence the LLR $L(y) := \log(f_1(y)/f_0(y))$ is linear as a function of y.
- The change time τ_{a} was a mixture of geometrics, hence satisfying (5). Two choices of distribution were used in this experiment, with $\varrho_{\mathsf{a}} = 0.02$ in each.
- Three choices for F were considered, each expressed as a log-likelihood ratio $F = \widehat{L} := \log(\check{f}_1/\check{f}_0)$, for which plots are shown in Fig. 2. The three plots in Fig. 3 show the evolution of $\{\mathcal{X}_n\}$ for each of three choices of F.

It was found in both ideal and mismatched settings that the error in the approximations given in Prop. 2.3 are nearly constant. That is, both $\bar{\mathrm{H}}^*(\kappa) - \bar{\mathrm{H}}_\infty^*(\kappa)$ and

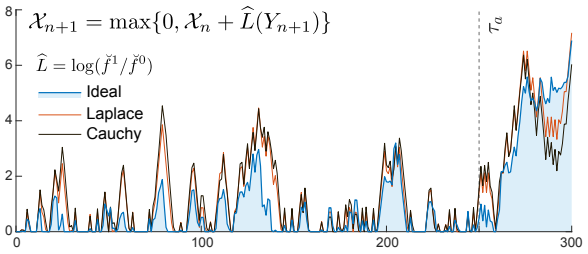

Fig. 3. Behavior of $\{\mathcal{X}_n\}$ for each of three choices of F in Fig. 2.

$\bar{J}^*(\kappa) - \bar{J}_\infty^*(\kappa)$ were nearly constant over the range considered, $2 \le \kappa \le 100$. This is evident in Fig. 4, in which the approximations in (17) were each shifted by a scalar constant to obtain an exact match at $\kappa = 100$.

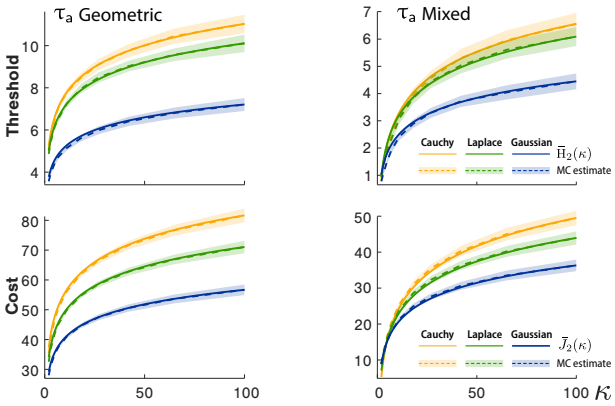

Fig. 4. Comparison of optimizing CUSUM threshold and policy with their approximations.

As discussed after Prop. 3.4, for this conditionally i.i.d. model the function $F^* = L + \varrho_{\mathsf{a}} := \log(f_1/f_0) + \varrho_{\mathsf{a}}$ is a minimizer of $\bar{J}_\infty^*(\kappa)$ over all F. We did not include the additive term ϱ_{a} in the previous experiments because we were not aware that F^* was a minimizer at the time the experiments were conducted.

We turn next to results from more recent experiments designed to test optimality of the additive constant. A scalar linear Gaussian Markov model was considered

$$\Phi_{k+1}^i = A^i \Phi_k^i + W_{k+1}^i + m_W^i$$

in which $|A^i| < 1$ for each i, and $\{W_k^i : k \ge 1\}$ is zero-mean, i.i.d. and Gaussian. For Markov models the function minimizing $\bar{J}_\infty^*(\kappa)$ depends on pairs of observations: $F^* = L_\infty + \varrho_{\mathsf{a}}$ with $X_k^i = (\Phi_{k-1}^i; \Phi_k^i)$. Based on the formula (6) we conclude that $F^* : \mathbb{R}^2 \to \mathbb{R}$ is quadratic.

Consider the special case in which both A^i and $\{W_k^i : k \ge 1\}$ do not depend upon i; that is, only the mean m_W^i changes at time τ_{a}. Simple calculations establish that $L_\infty(x, z) = a[z - Ax] + b$ for constants a, b, so that $\{L_\infty(\Phi_{k-1}^i, \Phi_k^i) : k \ge 1\}$ is i.i.d. for $i = 0$ or $i = 1$.

Consider next a change in dynamics only, in which $m_W^0 = m_W^1 = 0$, and each W_k^i is $N(0,1)$. In this case, eq. (6) is the quadratic

$$L_\infty(x,z) = \alpha x^2 - \beta xz\,,$$

with $\alpha = ([A^0]^2 - [A^1]^2)/2$ and $\beta = A^0 - A^1$. The numerics that follow used $A^0 = 0.4$ and $A^1 = 0.1$, giving $\alpha = 0.075$ and $\beta = 0.3$.

The distribution of the change time was taken to be geometric with parameter $\varrho_{\mathsf{a}} = 0.02$.

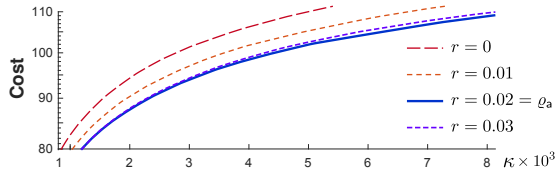

Fig. 5. Cost estimates obtained using Monte-Carlo for CUSUM using $F = L_\infty + r$ for four values of r, with L_∞ defined in (6).

Experiments were conducted to compare performance of CUSUM using various choices of the scalar r in $F = L_\infty + r$, including $r^* = \varrho_{\mathsf{a}} = 0.02$ which is predicted to be optimal for large κ. This is confirmed in the plots shown in Fig. 5. For the four values of r tested it was found that $r = 0.03$ yielded the smallest cost for $\kappa \le 1.3 \times 10^3$, while $r = \varrho_{\mathsf{a}}$ was best for $\kappa > 1.3 \times 10^3$,

IV. Conclusions

This work is intended to be a starting point for consideration of highly non-ideal settings faced in practice. In applications of interest to us there may be well understood behavior before a change (which might represent a fault in a transmission line, or a computer attack), bug little prior knowledge of post-change behavior.

Current research concerns relaxations of the independence assumption on τ_{a} and $\boldsymbol{X}^0$; preliminary theory is contained in [2]. Design of decision rules that are robust to post-change behavior might be inspired by the rich literature on robust hypothesis testing (see [12], [6] and the references therein).

References

[1] V. Anantharam. How large delays build up in a $GI/G/1$ queue. *Queueing Systems Theory Appl.*, 5(4):345–367, 1989.

[2] A. Cooper and S. Meyn. Quickest change detection using mismatched CUSUM. *arXiv 2409.07948*, 2024.

[3] A. Cooper and S. Meyn. Reinforcement learning design for quickest change detection. *IEEE CDC (to appear)*, 2024.

[4] A. Cooper and S. Meyn. Reinforcement learning design for quickest change detection—extended paper. *arXiv preprint arXiv:2403.14109*, 2024.

[5] A. Ganesh and N. O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastic Reports*, 73(1-2):25–35, 2002.

[6] D. Huang and S. Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Trans. Inform. Theory*, 59(12):8157–8181, 2013.

[7] I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13:304–362, 2003.

[8] I. Kontoyiannis and S. P. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, 10(3):61–123 (electronic), 2005.

[9] Y. Liang, A. G. Tartakovsky, and V. V. Veeravalli. Quickest change detection with non-stationary post-change observations. *arXiv 2110.01581*, 2021.

[10] Y. Liang and V. V. Veeravalli. Non-parametric quickest mean-change detection. *Transactions on Information Theory*, pages 8040–8052, 2022.

[11] A. N. Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007 (reprint from 1977 ed.).

[12] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli. Universal and composite hypothesis testing via mismatched divergence. *IEEE Trans. Inform. Theory*, 57(3):1587–1603, 2011.

[13] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

[14] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou. Data-driven quickest change detection in hidden Markov models. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2643–2648, June 2023.

[15] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou. Data-driven quickest change detection in Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023.

Appendix

A. Outline of proof of Prop. 2.3.

The first term in (2) is approximated by

$$\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_+] = \mathrm{H}/m_1 + O(1) \tag{26}$$

in which the error $O(1)$ is bounded as $\mathrm{H} \to \infty$.

The hard work is approximating the cost of eagerness, $\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-]$ *What is missing in this outline is justification for many of the approximations that follow.*

Eagerness and large deviations asymptotics. Independence of τ_{a}, $\boldsymbol{X}^0$ imply, for $k, n \ge 1$,

$$\mathsf{P}\{\tau_{\mathsf{s}} \le n \mid \tau_{\mathsf{a}} = n + k\} = \mathsf{P}\{\max_{0 \le t \le s_n} x_t^{(\mathsf{H})} \ge 1\}$$

where $s_n = n/\mathrm{H}$. Hence,

$$\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-] = \sum_{n=0}^{\infty} \mathsf{P}\{\max_{0 \le t \le s_n} x_t^{(\mathsf{H})} \ge 1\}\mathsf{P}\{\tau_{\mathsf{a}} > n\}$$

and applying Lemma 2.2 and (A2),

$$\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-] \approx \sum_{n=0}^{\infty} \exp\big(-\mathrm{H}G(s_n)\big) \tag{27}$$

in which $G(s_n) = e_0(s_n) + \varrho_{\mathsf{a}} s_n$.

Properties of G. It is convex on $\mathbb{R}_+$, with unique minimizer $s^* = 1/\breve{m}_+$ satisfying $G(s^*) = \mathfrak{v}_+$.

We have the Taylor series approximation,

$$\widetilde{G}(s) := G(s) - G(s^*) = \tfrac{1}{2}(s - s^*)^2/\gamma^2 + O(s - s^*)^3$$

where $\gamma^2 = 1/G''(s^*) = \Lambda_0''(\mathfrak{v}_+)/\mathfrak{v}_+^3$.

Integral approximation. Let $n^* = \lfloor \mathrm{H}/\breve{m}_+ \rfloor$.

Setting $\delta = \mathrm{H}^{-1+\varepsilon}$ with $\varepsilon \in (1/2, 1)$, the following approximations may be justified:

$$\begin{aligned}\sum_{n=0}^{\infty} \exp\bigl(-\mathrm{H}\widetilde{G}(s_n)\bigr) &\approx \sum_{|n-n^*|\le\delta} \exp\bigl(-\mathrm{H}\widetilde{G}(s_n)\bigr) \\ &\approx \mathrm{H}\int_{|s-s^*|\le\delta} \exp\bigl(-\mathrm{H}\widetilde{G}(s)\bigr)\,ds \\ &\approx \mathrm{H}\int_{-\infty}^{\infty} \exp\bigl(-\tfrac{(s-s^*)^2}{2\sigma^2}\bigr)\,ds\end{aligned}$$

with $\sigma^2 = \gamma^2/\mathrm{H}$. Combining this approximation with (27) gives $\mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_-] \approx \sqrt{\mathrm{H}}\sqrt{2\pi\gamma^2}\exp(-\mathrm{H}G(s^*))$. In view of (26) and the identity $G(s^*) = \upsilon_+$,

$$\bar{J}(\mathrm{H}, \kappa) \approx \mathrm{H}/m_1 + \kappa\sqrt{\mathrm{H}}\sqrt{2\pi\gamma^2}\exp(-\mathrm{H}\upsilon_+) \qquad (28)$$

Optimizing the approximation. An exact minimum of the RHS of (28) is not available, but approximations can be found based on the first order condition for optimality, which admits the approximation

$$\sqrt{\mathrm{H}^*}\sqrt{2\pi\gamma^2}\exp(-\mathrm{H}^*\upsilon_+) \approx \frac{1}{m_1\upsilon_+}\frac{1}{\kappa} \qquad (29)$$

Further calculations lead to (18a).

From (28) we have $\bar{J}^*(\kappa) = \bar{J}(\mathrm{H}^*, \kappa) + o(\log(\kappa))$, for which an approximation requires examination of two terms,

$$\begin{aligned}&\bar{J}^*(\kappa) = \bar{J}_0(\kappa) + \bar{J}_1(\kappa) + o(\log(\kappa)) \\ &\quad \textit{with} \quad \bar{J}_0(\kappa) = \mathrm{H}^*/m_1 \\ &\qquad\qquad \bar{J}_1(\kappa) = \kappa\sqrt{\mathrm{H}^*}\sqrt{2\pi\gamma^2}\exp(-\mathrm{H}^*\upsilon_+)\end{aligned}$$

The term $\bar{J}_1(\kappa)$ is bounded in κ due to (29).

Also from (29), exactly as in the derivation of (18a), we have $\bar{J}_0(\kappa) = \bar{\mathrm{H}}^*_\infty(\kappa)/m_1 + o(\log(\kappa))$ as required.

■