IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Anomaly Search Over Many Sequences With Switching Costs

Matthew Ubl, *Member, IEEE*, Benjamin D. Robinson, *Member, IEEE*, and Matthew Hale, *Member, IEEE*

***Abstract*—This paper considers the quickest search problem to identify anomalies among large numbers of data streams. These streams can model, for example, disjoint regions monitored by a mobile robot. A particular challenge is a version of the problem in which the experimenter must suffer a cost each time the data stream being sampled changes, such as the time the robot must spend moving between regions. In this paper, we propose an algorithm which accounts for switching costs by varying a confidence threshold that governs when the algorithm switches to a new data stream. Our main contributions are easily computable approximations for both the optimal value of this threshold and the optimal value of the parameter that determines when a stream is flagged as an anomaly, using the Brownian motion approximations. Further, we empirically show (i) a uniform improvement for switching costs of interest and (ii) roughly equivalent performance for small switching costs when comparing to the closest available algorithm.**

***Index Terms*—Quickest search, sequential analysis, controlled sensing, scanning rule, switching costs**

I. Introduction

A fundamental problem in sensing and signal processing is online anomaly search. With origins in Chernoff's sequential design of experiments [1], the aim of online anomaly search is to develop an efficient policy for sampling a subset of several data streams over time that quickly and accurately identifies any anomalous ones. Some variations on Chernoff's method include [2], [3], [4], [5], [6], [7], [8]. An important application is clinical-trial design, where the drug tested in each trial depends on the drugs tested and outcomes obtained in all previous trials of the study. Another application is online search for an open channel in cognitive radio, where the sampling policy may depend on past observations.

Recently several authors have considered the problem of online anomaly search in which a cost is incurred any time the current channel sampled differs from the last one [9], [10], [11]. This new formulation models online anomaly search more realistically than the traditional formulation because it accounts for possible switching costs in hardware or software, e.g., if an autonomous agent must move to observe a new location or if equipment must be repositioned to do so. This formulation is also more challenging because conventional analyses of exploration versus exploitation do not apply, and thus many existing works do not consider it [12]. Table 1 shows how the problem setup in this paper relates to that of existing work. We emphasize that we differ from these existing works by considering switching costs and an infinite number of data streams simultaneously. To the best of our knowledge, this is the first work to do so. Furthermore, these existing works that address switching costs for finite data streams use methods that are either not possible to implement for infinitely many data streams or provide no tangible benefit in a setting with infinitely many data streams, such as round-robin sampling, revisiting previous data streams, or block-scheduling observations [11], [13], [14], [15].

Sequences	Switching costs	No switching costs
Finite	[11], [13], [14], [15]	[1] [2]
Infinite	This work	[12]

Fig. 1. Relationship of this work (lower left quadrant) to existing works; the other three quadrants contain references to representative existing works on those classes of problems.

Our approach is to adapt an existing method to the case of switching costs. We do this by (i) introducing a new parameter that controls switching between data streams and (ii) numerically optimizing both this switching parameter and an existing stopping parameter. Our contributions are:

- We develop an algorithm that generalizes the algorithm in [12], which is known to be Bayes-optimal for the setting without switching costs.
- We show that in a certain asymptotic regime, the optimal parameters for this algorithm can be found by exactly solving an algebraic equation and numerically solving a strongly convex optimization problem over a scalar decision variable. This approach compares favorably with the Monte Carlo method of [12] in terms of computational efficiency.
- We illustrate an almost uniform improvement in performance over the closest existing method in experiments. We illustrate empirically a uniform improvement for switching costs of interest and roughly equivalent performance for small switching costs, compared to [12].

The rest of this paper is organized as follows. In Section II we provide our problem setup and discuss the existing optimal solution for the setting without switching costs. In Section III we introduce switching costs and derive an approximation of the combined observation-switching cost, and derive the parameter choices that minimize this approximation.

Date submitted: 2 March 2023. AFRL Public Affair number AFRL-2023-0996. This work was supported by the Office of Naval Research under grants N00014-22-1-2435 and N00014-21-1-2495, by the Air Force Office of Scientific Research under grants FA9550-19-1-0169 and RYCOR036, and by the Air Force Research Lab Sensors Directorate.

Benjamin Robinson is with the Air Force Research Lab Sensors Directorate, WPAFB, OH 45433 USA. (Email: benjamin.robinson.8@us.af.mil)

Matthew Ubl and Matthew Hale are with the University of Florida Aerospace Engineering Department, Gainesville, FL, 32611 USA. (Respective emails: m.ubl@ufl.edu and matthewhale@ufl.edu)

In Section IV we justify this approximation and explore the conditions for which it is accurate. In Section V we provide numerical results, and we conclude in Section VI.

II. Preliminaries

A. Problem Setup

We consider data streams indexed over $k \in \{1, 2, \dots\}$. Data stream k generates the i.i.d. sequence of random variables $X_1^k, X_2^k, \dots$ which have sample space $\mathcal{X}$. Each data stream obeys one of two hypotheses, H_0 or H_1. Consider two distinct distributions on $\mathcal{X}$: F_0 and F_1. We say hypothesis H_0 is true for data stream k if $X_t^k \sim F_0$ for all $t \in \{1, 2, \dots\}$, and that H_1 is true for data stream k if $X_t^k \sim F_1$ for all $t \in \{1, 2, \dots\}$. We use f_0 and f_1 to denote the PDFs of F_0 and F_1 respectively. In this paper if H_0 is true for a particular data stream, we say that stream is *nominal*, and if H_1 is true, that the stream is a *target* stream. We assume that for any given data stream, H_1 is true with prior probability $\hat{\pi}$ and H_0 is true with prior probability $1 - \hat{\pi}$, where $\hat{\pi} \in (0, 1)$. We emphasize that we make no assumptions regarding the value of $\hat{\pi}$, i.e., we do not assume or require that target streams are "rare". We use $\mathbb{E}_i$ to denote the expectation under hypothesis i, and define $\mathbb{E}_{\hat{\pi}}[\,\cdot\,] = \hat{\pi}\mathbb{E}_1[\,\cdot\,] + (1 - \hat{\pi})\mathbb{E}_0[\,\cdot\,]$.

We assume we have a single observer that can sample one and only one data stream at a time. That is, if the observer samples data stream k at time t it receives X_t^k but no information from the other data streams. Our goal is to design an algorithm for this observer to identify a target data stream as quickly as possible. In much of the existing literature, "as quickly as possible" means minimizing the expected number of observations required while satisfying some constraint on the error probability, i.e., satisfying an upper bound on the probability that the stream we identify as a target is actually nominal. We use τ to denote the number of observations taken before the algorithm terminates and declares a particular data stream, which we denote as k_τ, as a target. We use H^{k_τ} to denote the hypothesis obeyed by data stream k_τ, and $P(H^{k_\tau} = H_0)$ as the error rate, i.e., the probability that the stream k_τ is actually nominal. Therefore, our goal is to find the algorithm that minimizes $\mathbb{E}_{\hat{\pi}}[\tau]$ while ensuring $P(H^{k_\tau} = H_0) \leq \epsilon$, where $\epsilon > 0$ is an allowable error rate.

B. Solution Without Switching Costs

This subsection briefly reviews related work on problems without switching costs; switching costs will be introduced in the next section. It was shown in [12] that a cumulative-sum-based (CUSUM-based) test is the optimal algorithm for the setting without switching costs. This algorithm is defined by two threshold parameters: $\gamma_L \leq 0 \leq \gamma_U$. In this algorithm the observer maintains a statistic Λ_t^k for stream k which is updated after every observation and is initialized as $\Lambda_0^1 = 0$. If at time t the observer samples data stream k (i.e., the observer receives X_t^k), then it performs the update $\Lambda_t^k = \Lambda_{t-1}^k + \log\left(\frac{f_1(X_t^k)}{f_0(X_t^k)}\right)$. If $\gamma_L \leq \Lambda_t^k < \gamma_U$, then the observer will sample data stream k again at time $t + 1$. If $\Lambda_t^k < \gamma_L$, then the observer has declared stream k as nominal. The observer will begin using the new statistic $\Lambda_t^{k+1} = 0$ and will switch to sampling data stream $k + 1$ (we assume the streams are either pre-ordered or the next stream is selected at random) beginning at time $t+1$. This procedure describes the k^{th} *stage* of the algorithm, during which data stream k is observed.

The algorithm carries out the same procedure on data streams $k + 1, k + 2, \dots$ until a target stream is indicated. Specifically, if $\Lambda_t^k \geq \gamma_U$, then the algorithm terminates and the observer declares data stream k as a target stream. The optimal choice for γ_L is 0 regardless of F_0, F_1, $\hat{\pi}$, or ϵ [12, Section IV]. Because $\mathbb{E}_{\hat{\pi}}[\tau]$ monotonically increases as $\gamma_U \to \infty$ and $P(H^{k_\tau} = H_0)$ monotonically decreases as $\gamma_U \to \infty$, the optimal choice for γ_U is the smallest value for which $P(H^{k_\tau} = H_0) \leq \epsilon$. However, a closed form for this γ_U is not known; it must be estimated using numerical experiments.

III. Problem Statement with Switching Costs

We now introduce switching costs to the model described in the previous section, and this will give the problem formulation that we consider in this paper. When the observer switches from data stream k to data stream $k + 1$, it now incurs a cost λ_k drawn from some non-negative distribution L for all k. That is, $\lambda_k \geq 0$ and $\mathbb{E}[\lambda_k] = \bar{\lambda}$ is finite. We also assume that the costs λ_k and the observations $X_t^{k'}$ are mutually independent for all k, k', and t. This cost models applications in which observing a new data stream requires "deadtime" when no observations can be taken, such as when equipment needs to be re-positioned or re-calibrated. It also models problems in which observations and switches are "costly" in some resource other than time, such as energy or money. Under this switching cost assumption, the new optimization problem becomes:

Problem 1. Let an error tolerance $\epsilon \in (0, 1)$ and prior $\hat{\pi} \in (0, 1)$ be given. Then

$$\text{minimize}_{\gamma_L \leq 0 \leq \gamma_U} \mathbb{E}_{\hat{\pi}}[\tau] + \mathbb{E}_{\hat{\pi}}[s] \tag{1}$$

$$\text{s.t. } P(H^{k_\tau} = H_0) \leq \epsilon, \tag{2}$$

where $s = \sum_{i=1}^{k_\tau - 1} \lambda_k$ is the total switching cost incurred before terminating the algorithm.

To derive threshold choices for this problem we will next rewrite the problem in terms of the *stage-wise* false-positive and false-negative rates α and β of the algorithm, and then establish the relationship between these rates and the threshold choices γ_L and γ_U. By the stage-wise false-positive rate, we mean the probability that a stage's terminal value of Λ exceeds (or equals) γ_U given that the stage's data follow H_0, i.e., the probability that a stream is declared a target even though the stage's data are nominal. The stage-wise false-negative rate is similarly defined as the probability that γ_L exceeds the stage's terminal value of Λ given that the stage's data follow H_1, i.e., the stage's data are not nominal, but the stream is labelled nominal.

Let t_k be the time when the algorithm takes its last observation of stream k (i.e., $\Lambda_{t_k}^k \notin [\gamma_L, \gamma_U)$). Then $\hat{t} = t_k - t_{k-1}$ is the *stage-wise stopping time* of stage k, or the number of observations taken of stream k before making a decision. Then we may more compactly write that $\alpha = \mathbb{P}_0[\Lambda_{t_k}^k \geq \gamma_U]$

and $\beta = \mathbb{P}_1[\Lambda^k_{t_k} < \gamma_L]$, where $\mathbb{P}_i$ is the probability under hypothesis i. Consider the inequalities

$$\gamma_L \geq \log(\beta/(1-\alpha)) \tag{3}$$
$$\gamma_U \leq \log((1-\beta)/\alpha)) \tag{4}$$
$$\mathbb{E}_{\hat{\pi}}[\Lambda^k_{t_k} \mid \Lambda^k_{t_k} \geq \gamma_U] \geq \gamma_U \tag{5}$$
$$\mathbb{E}_{\hat{\pi}}[\Lambda^k_{t_k} \mid \Lambda^k_{t_k} < \gamma_L] \leq \gamma_L, \tag{6}$$

which are given in [16, Equations (2.9) and (2.10)]. In accordance with [16], we assume these inequalites are approximate equalities for the remainder of this section. These approximations are known as "Brownian motion approximations".

Remark 1 (Brownian Motion Approximations)**.** The assumption that (3)-(6) are approximate equalities is accurate under the conditions that (a) f_0 and f_1 are sufficiently close, (b) $\bar{\lambda}$ is sufficiently large, and (c) ϵ is small. These conditions imply that the step size in the sequential probability ratio test of one stage is small compared to the decision thresholds, and thus overshoots of decision thresholds are also comparatively small. We will elaborate on these conditions in Section IV.

The quantities α and β appear in Problem 1 in the following manner: using Wald's Identity we see that $\mathbb{E}_{\hat{\pi}}[\tau] = \mathbb{E}_{\hat{\pi}}[k_\tau]\mathbb{E}_{\hat{\pi}}[\hat{t}]$ (see Equation (30) in [12]). Intuitively, this result states that the expected number of total observations before termination is equal to the expected number of data streams visited (k_τ) multiplied by the expected number of observations per data stream (which is $\hat{t}$). Using Wald's Identity again gives $\mathbb{E}_{\hat{\pi}}[s] = \mathbb{E}_{\hat{\pi}}[k_\tau - 1]\bar{\lambda}$, which states that the expected total switching cost incurred over time is equal to the expected number of switches (one less than the number of streams visited) multiplied by the expected switching cost per switch.

We first address the term $\mathbb{E}_{\hat{\pi}}[\hat{t}]$. Assume that a particular data stream k being sampled by the observer is a target, and assume that the observer takes its last sample of k at time t_k. Then one of the termination criteria has been met and $\Lambda^k_{t_k} \notin [\gamma_L, \gamma_U)$. Let f_0, f_1 be the PDFs of F_0, F_1 respectively. Using Wald's Identity again, we have $\mathbb{E}_1[\Lambda^k_{t_k}] = D(f_1||f_0)\mathbb{E}_1[t_k]$, where $D(f_1||f_0) = \mathbb{E}_1\left[\log \frac{f_1(x)}{f_0(x)}\right]$ is the Kullback-Leibler (KL) divergence of f_0 from f_1. Using the same procedure we see $\mathbb{E}_0[\Lambda^k_{t_k}] = -D(f_0||f_1)\mathbb{E}_0[\hat{t}]$. From the definition of $\mathbb{E}_{\hat{\pi}}[\cdot]$, we can write

$$\mathbb{E}_{\hat{\pi}}[\hat{t}] = (1-\hat{\pi})\frac{\mathbb{E}_0[\Lambda^k_{t_k}]}{-D(f_0||f_1)} + \hat{\pi}\frac{\mathbb{E}_1[\Lambda^k_{t_k}]}{D(f_1||f_0)}. \tag{7}$$

Furthermore, we see that $\mathbb{E}_0[\Lambda^k_{t_k}] = \alpha\mathbb{E}_0[\Lambda^k_{t_k}|\Lambda^k_{t_k} \geq \gamma_U] + (1-\alpha)\mathbb{E}_0[\Lambda^k_{t_k}|\Lambda^k_{t_k} < \gamma_L]$. By the Brownian motion approximations we have $\mathbb{E}_0[\Lambda^k_{t_k}|\Lambda^k_{t_k} \geq \gamma_U] \approx \gamma_U$ and $\mathbb{E}_0[\Lambda^k_{t_k}|\Lambda^k_{t_k} < \gamma_L] \approx \gamma_L$. Following equivalent steps for $\mathbb{E}_1[\Lambda^k_{t_k}]$ gives

$$\mathbb{E}_{\hat{\pi}}[\hat{t}] \approx (1-\hat{\pi})\frac{\alpha\gamma_U + (1-\alpha)\gamma_L}{-D(f_0||f_1)} + \hat{\pi}\frac{(1-\beta)\gamma_U + \beta\gamma_L}{D(f_1||f_0)}. \tag{8}$$

Furthermore, from [12, Equation (30)] we also have

$$\mathbb{E}_{\hat{\pi}}[k_\tau] = \frac{1}{(1-\hat{\pi})\alpha + \hat{\pi}(1-\beta)}. \tag{9}$$

For ease of notation, we now define $\delta_U = \exp(\gamma_U)$ and $\delta_L = \exp(\gamma_L)$. While $\gamma_L \leq 0 \leq \gamma_U$ are the actual thresholds used by the algorithm, rewriting the problem in terms of $0 < \delta_L \leq 1 \leq \delta_U$ makes the following notation simpler. Inverting the approximate equalities in (3) and (4), we obtain the following: $\alpha \approx \frac{1-\delta_L}{\delta_U - \delta_L}$, $\beta \approx \delta_L\frac{\delta_U - 1}{\delta_U - \delta_L}$, $1-\alpha \approx \frac{\delta_U - 1}{\delta_U - \delta_L}$, and $1-\beta \approx \delta_U\frac{1-\delta_L}{\delta_U - \delta_L}$. Substituting these into (8) and (9) and simplifying yields that the function

$$\begin{aligned} C(\delta_L, \delta_U) := \frac{1-\hat{\pi}}{-D(f_0||f_1)}\frac{\log(\delta_U) + \frac{\delta_U - 1}{1-\delta_L}\log(\delta_L)}{1 + \hat{\pi}(\delta_U - 1)} \\ + \frac{\hat{\pi}}{D(f_1||f_0)}\frac{\delta_U\log(\delta_U) + \delta_L\frac{\delta_U - 1}{1-\delta_L}\log(\delta_L)}{1 + \hat{\pi}(\delta_U - 1)} \\ + \frac{\bar{\lambda}\frac{\delta_U - \delta_L}{1-\delta_L}}{1 + \hat{\pi}(\delta_U - 1)} \end{aligned} \tag{10}$$

approximates the cost given in (1).

Furthermore, from [12, Equation (29)] we have that that $P(H^{k_\tau} = H_0) = \frac{(1-\hat{\pi})\alpha}{(1-\hat{\pi})\alpha + \hat{\pi}(1-\beta)}$, which is bounded above by $\frac{1-\hat{\pi}}{1+\hat{\pi}(\delta_U - 1)}$ since $(1-\beta)/\alpha \geq \delta_U$. As a result, this inequality is an approximate equality under the Brownian motion approximations. Therefore, following some algebraic manipulation we see that $P(H^{k_\tau} = H_0) \leq \epsilon$ if $\delta_U \geq \frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon}$. Note that $\frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon} > 1$ so long as $\epsilon < 1 - \hat{\pi}$, which should always be the case; if the tolerable error rate is greater than the prevalence of nominal data streams then the optimal algorithm is to take no observations and flag a data stream at random.

Therefore, we can find an approximate solution to Problem 1 by solving the following problem.

Problem 2. Let an error tolerance $\epsilon \in (0,1)$ and a prior $\hat{\pi} \in (0,1)$ be given. Then find

$$\text{Find } (\delta_L^*, \delta_U^*) = \underset{\delta_L, \delta_U}{\arg\min}\ C(\delta_L, \delta_U) \tag{11}$$
$$\text{s.t. } \delta_U \geq \frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon} \tag{12}$$
$$\delta_L \in [0,1]. \tag{13}$$

From the structure of $C(\delta_L, \delta_U)$, we can derive the following propositions:

Proposition 1. *For any fixed $\hat{\delta}_L \in (0,1)$, $C(\hat{\delta}_L, \cdot)$ is monotonically increasing on the domain $[1, \infty)$. From this fact, we get $\delta_U^* = \frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon}$.*

Proof: The monotonic behavior of $C(\hat{\delta}_L, \cdot)$ on this domain is apparent by inspection. Because we wish to minimize $C(\hat{\delta}_L, \cdot)$, we want to set δ_U to its minimum allowable value. Since that value is $\frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon}$, we have $\delta_U^* = \frac{1-\hat{\pi}}{\hat{\pi}}\frac{1-\epsilon}{\epsilon}$ regardless of the value of $\hat{\delta}_L$. ■

Proposition 2. *For any fixed $\hat{\delta}_U > 1$, $C(\cdot, \hat{\delta}_U)$ is strongly convex on the domain $(0,1]$. From this fact δ_L^* exists, is unique, and can be found by solving the scalar optimization problem $\delta_L^* = \underset{\delta_L \in [0,1]}{\arg\min}\ C(\delta_L, \delta_U^*)$. Furthermore, δ_L^* lies in the interior of this interval for $\bar{\lambda} > 0$ (i.e., $\delta_L^* \in (0,1)$).*

Proof: Differentiation shows $\lim_{\delta_L \to 0^+}\frac{\partial C(\delta_L, \hat{\delta}_U)}{\partial \delta_L} = -\infty$ and $\lim_{\delta_L \to 1^-}\frac{\partial C(\delta_L, \hat{\delta}_U)}{\partial \delta_L} = \infty$ so long as $\bar{\lambda} > 0$. Therefore, from the Intermediate Value Theorem there must exist some

Algorithm 1: Quickest Search Algorithm with Switching Costs

Input: $\epsilon \in (0,1)$, $\hat{\pi} \in (0,1)$, $D(f_1||f_0) > 0$, $D(f_0||f_1) > 0$
$\gamma_U \leftarrow \log\left(\frac{1-\epsilon}{\epsilon}\frac{1-\hat{\pi}}{\hat{\pi}}\right)$
$\gamma_L \leftarrow \log\left(\arg\min_{[0,1]} C\left(\cdot, \frac{1-\epsilon}{\epsilon}\frac{1-\hat{\pi}}{\hat{\pi}}\right)\right)$
$t \leftarrow 0$, $\Lambda_0^1 \leftarrow 0$, $k \leftarrow 1$
while $\Lambda_t^k < \gamma_U$ **do**
  if $\Lambda_t^k \geq \gamma_L$ **then**
    Observe: X_{t+1}^k
    $\Lambda_{t+1}^k \leftarrow \Lambda_t^k + \log\left(\frac{f_1(X_{t+1}^k)}{f_0(X_{t+1}^k)}\right)$
    $t \leftarrow t+1$
  else
    $k \leftarrow k+1$
    $\Lambda_t^k \leftarrow 0$
  end if
end while
Label arm k as a target

value $\hat{\delta}_L \in (0,1)$ for which $\frac{\partial C(\delta_L, \hat{\delta}_U)}{\partial \delta_L} = 0$ at $\delta_L = \hat{\delta}_L$. Strong convexity is established by characterizing the limiting behavior of $\frac{\partial^2 C(\delta_L, \hat{\delta}_U)}{\partial \delta_L^2}$. Observe that $C(\delta_L, \delta_U)$ is a sum of three terms: the first two which contain the KL divergences $D(f_0||f_1)$ and $D(f_1||f_0)$, and the third which contains $\bar{\lambda}$. Name these terms C_1 C_2, and C_3 respectively. First, we see that $\lim_{\delta_L \to 0^+} \frac{\partial^2 (C_1(\delta_L, \hat{\delta}_U) + C_2(\delta_L, \hat{\delta}_U))}{\partial \delta_L^2} = \infty$ and $\lim_{\delta_L \to 1^-} \frac{\partial^2 (C_1(\delta_L, \hat{\delta}_U) + C_2(\delta_L, \hat{\delta}_U))}{\partial \delta_L^2} = \frac{\hat{\delta}_U - 1}{1 + \hat{\pi}(\hat{\delta}_U - 1)}\left(\frac{2}{3}\frac{1-\hat{\pi}}{D(f_0||f_1)} + \frac{1}{3}\frac{\hat{\pi}}{D(f_1||f_0)}\right)$, and that $\frac{\partial^2 (C_1(\delta_L, \hat{\delta}_U) + C_2(\delta_L, \hat{\delta}_U))}{\partial \delta_L^2}$ is monotonically decreasing with δ_L on $[0,1]$. Additionally, we see $\lim_{\delta_L \to 0^+} \frac{\partial^2 C_3(\delta_L, \hat{\delta}_U)}{\partial \delta_L^2} = \frac{\bar{\lambda}(\hat{\delta}_U - 1)}{1 + \hat{\pi}(\hat{\delta}_U - 1)}$ and $\lim_{\delta_L \to 1^-} \frac{\partial^2 C_3(\delta_L, \hat{\delta}_U)}{\partial \delta_L^2} = \infty$, and that $\frac{\partial^2 C_3(\delta_L, \hat{\delta}_U)}{\partial \delta_L^2}$ is monotonically increasing with δ_L on $[0,1]$. Therefore, $C(\cdot, \hat{\delta}_U)$ is $\frac{\hat{\delta}_U - 1}{1 + \hat{\pi}(\hat{\delta}_U - 1)}\left(\frac{2}{3}\frac{1-\hat{\pi}}{D(f_0||f_1)} + \frac{1}{3}\frac{\hat{\pi}}{D(f_1||f_0)} + \bar{\lambda}\right)$-strongly convex on this interval. Therefore $\hat{\delta}_L$ is a unique minimizer of $C(\cdot, \hat{\delta}_U)$, and δ_L^* can be found by minimizing $C(\cdot, \delta_U^*)$. ■

Therefore, Propositions 1 and 2 tell us we can calculate δ_U^* explicitly as a function of $\hat{\pi}$ and ϵ, and δ_L^* numerically as the solution to a scalar, set-constrained, strongly convex optimization problem. These steps give rise to the Quickest Search Algorithm with switching Costs, which is Algorithm 1.

IV. Discussion of Brownian-Motion Approximations

Now that we have presented Problem 2 as a solvable approximation of Problem 1, we will justify this substitution by showing that in the limiting cases described in Remark 1 in Section III, the approximate inequalities used to formulate Problem 2 approach equalities. We are interested in the case where f_0 and f_1 are "close" since if they are easily distinguished, the problem is easy and optimality is not crucial. We are interested in the case where ϵ is small because we want few errors. Further, we are interested in the case where $\bar{\lambda}$ is relatively large because otherwise, the problem is solvable by the existing method of [12].

Recall from [16] that that the inequalities (3)-(6) being treated as equalities only fail to be equalities if the statistic $\Lambda_{t_k}^k$ overshoots the relevant threshold γ_U or γ_L, rather than hitting it exactly. That is, while we will always have $\Lambda_{t_k}^k \geq \gamma_U$ (or $\Lambda_{t_k}^k < \gamma_L$) at the end of any stage, Problem 2 is derived by assuming $\Lambda_{t_k}^k = \gamma_U$ (or $\Lambda_{t_k}^k = \gamma_L$). This approximation is reasonable when the expected overshoot of a particular threshold is small with respect to the threshold itself, i.e., if $\mathbb{E}\left[\frac{\Lambda_{t_k}^k - \gamma_U}{\gamma_U} \mid \Lambda_{t_k}^k \geq \gamma_U\right]$ and $\mathbb{E}\left[\frac{\Lambda_{t_k}^k - \gamma_L}{\gamma_L} \mid \Lambda_{t_k}^k < \gamma_L\right]$ are small. The remainder of this section shows that these terms are indeed small when a problem satisfies the conditions in Remark 1.

A. The case of "close" f_0 and f_1

Here the phrase "sufficiently close" means the KL divergences $D(f_1||f_0)$ and $D(f_0||f_1)$ are small. Because $\mathbb{E}[\Lambda_{t_k}^k - \gamma_U | \Lambda_{t_k}^k \geq \gamma_U] \leq \mathbb{E}[\Lambda_{t_k}^k - \Lambda_{t_k-1}^k | \Lambda_{t_k}^k \geq \gamma_U]$, we can see from the update law for Λ^k in Algorithm 1 and the definition of the KL divergences that this expected overshoot approaches zero as $D(f_1||f_0)$ and $D(f_0||f_1)$ approach zero, as desired.

B. The case of small ϵ

As ϵ shrinks, we must enforce a smaller error probability $P(H^{k_\tau} = H_0)$. From its definition, a smaller error probability directly implies a larger value of $\frac{1-\beta}{\alpha}$, which implies a larger value of γ_U^*. This relationship is intuitive: while both γ_U and γ_L affect $P(H^{k_\tau} = H_0)$, the effect of γ_U is significantly greater since the algorithm only terminates at data stream k if $\Lambda_{t_k}^k \geq \gamma_U$. Having a large γ_U^* means the expected overshoots are small, as desired.

C. The case of large $\bar{\lambda}$

The relationship between $\bar{\lambda}$ and γ_L^* is perhaps the most interesting one. Consider the high-level goal of our analysis: to minimize the number of switches our algorithm makes before finding and identifying (hopefully correctly) a target data stream. The requirement that we find a target data stream (with probability $1-\epsilon$) means that we specifically want to avoid switching away from a target data stream. In other words, the goal is to reduce β. As with $P(H^{k_\tau} = H_0)$, β depends on both γ_U and γ_L, but the effect of γ_L is significantly greater as the algorithm only switches away from stream k if $\Lambda_{t_k}^k < \gamma_L$. Having a very negative γ_L^* means the expected overshoots are close to zero, as desired.

The purpose of this analysis is to address non-trivial switching costs. However, we do note that our rule for selecting γ_L^* is optimal as $\bar{\lambda}$ approaches zero as well. Recall from Section II and [12] that the true optimal value of γ_L for $\bar{\lambda} = 0$ (i.e., the value that minimizes (1)) is $\gamma_L = 0$. As $\bar{\lambda} \to 0$, our value of γ_L^* found by solving Problem 2 also approaches zero, implying that the algorithm described in [12] is a special case of the one we develop here.

V. Numerical Results

We now compare the performance of our algorithm with the one described in [12] in MATLAB, which is the closest comparable algorithm, in a setting where switching costs are present. While the algorithm in [12] is optimal for the case where $\bar{\lambda} = 0$, it does not take into account switching costs. Furthermore, the optimal threshold γ_U cannot be directly calculated for that algorithm, and must be estimated via Monte Carlo simulations. In contrast, our algorithm directly accounts for switching costs and uses thresholds that can be directly calculated.

In this setting, target data streams occur with prior probability $\hat{\pi} = 0.1$, and obey the distribution $F_1 = \mathcal{N}(0, 1)$. Nominal data streams obey $F_0 = \mathcal{N}(0, 1.5)$. That is, for the purposes of this simulation, if an algorithm crosses γ_L, the next set of observations have a 10% chance of being drawn from f_1, and f_0 otherwise. We choose $\epsilon = 0.01$ to be our maximum tolerable error rate. Switching costs are drawn from a gamma distribution $\lambda_k \sim \Gamma(a, b)$, which has $\bar{\lambda} = \frac{a}{b}$. We will compare the performances of both algorithms across a range of values of $\bar{\lambda}$, by keeping $b = 1$ constant and exploring $a \in [0, 5]$. The algorithm from [12] uses thresholds $\gamma_L = 0$ and $\gamma_U = 6.130$ for this problem regardless of the switching costs. For the algorithm described in this paper, $\gamma_U = 6.794$ regardless of switching costs, and γ_L is chosen by solving Problem 2 with $\bar{\lambda}$. The ordering of nominal vs target data streams is randomly generated, but kept the same for both algorithms for a fair comparison and reproducibility. The values of γ_L used and their corresponding values of $\bar{\lambda}$ are plotted in Figure 2.

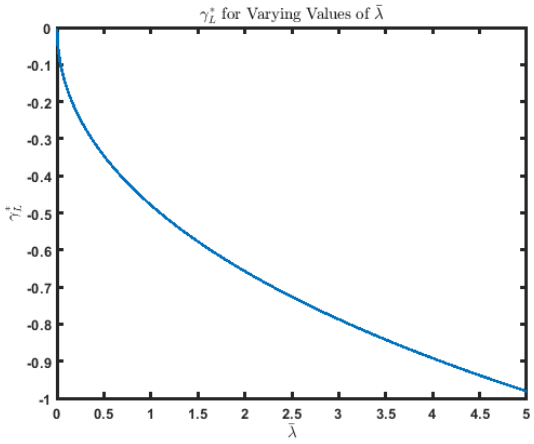

Fig. 2. The larger $\bar{\lambda}$ is, the more negative γ_L^* is in Algorithm 1.

The algorithm from [12] achieves $\mathbb{E}_{\hat{\pi}}[\tau] = 109.42$ and $\mathbb{E}_{\hat{\pi}}[k_\tau - 1] = 42.15$, and our algorithm achieves the similar numbers $\mathbb{E}_{\hat{\pi}}[\tau] = 113.21$ and $\mathbb{E}_{\hat{\pi}}[k_\tau - 1] = 42.04$ for $\bar{\lambda} = 0$. We note that in this zero switching cost case our algorithm achieves an error of 3.4% compared to the algorithm in [12], demonstrating that the Brownian Motion Approximations resulted in a valid approximation for this problem. Furthermore, from Section IV our approximated thresholds will more closely approach the theoretically optimal thresholds as $\bar{\lambda}$ grows. We also note that our algorithm achieves an error rate of 0.005, which satisfies our error bound of $\epsilon = 0.01$. As $\bar{\lambda}$ grows, our goal is to reduce the combined observation/switching cost formulated in (1) by reducing the number of switches. In Figure 3, we see that this is achieved. As γ_L is varied to account for higher values of $\bar{\lambda}$, we see that the number of expected switches drops significantly. Note that, given the prior $\hat{\pi} = 0.1$, a "perfect" algorithm (one that perfectly identifies all nominal and target streams and achieves $\alpha, \beta = 0$) would have $\mathbb{E}_{\hat{\pi}=0.1}[k_\tau - 1] = 9$, since on average the algorithm would have to scan through 10 streams before encountering its first target stream. As such, as $\bar{\lambda}$ grows and γ_L becomes more negative, we would expect the number of switches for our algorithm to approach 9, which is the behavior observed in Figure 3.

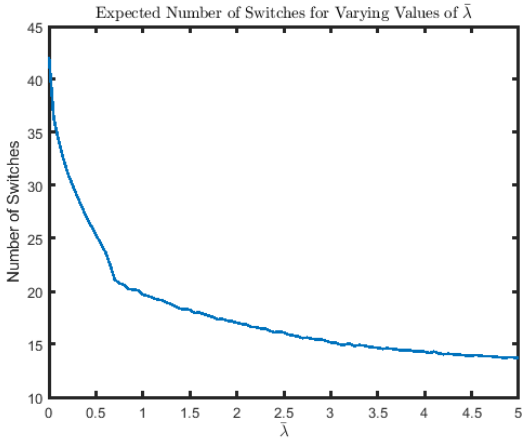

Fig. 3. The expected number of switches under Algorithm 1 for different values of $\bar{\lambda}$. As $\gamma_L \to -\infty$, the expected number of switches will approach 9, the expected number of switches required before our algorithm encounters its first target stream.

As a consequence of a more negative value of γ_L, our algorithm will tend to take more observations of data streams before switching away. The end result is, as Figure 4 shows, the total number of observations before termination of our algorithm grows with $\bar{\lambda}$. However, as we will see in Figure 5, this growth in observation cost is more than offset by the reduction in expected switching cost observed in Figure 3.

The combined observation/switching costs for both algorithms are plotted in Figure 5. We can see that in the large $\bar{\lambda}$ region, our algorithm significantly outperforms the algorithm from [12], which, due to its fixed behavior, achieves a cost of $109.42 + \bar{\lambda}42.15$, resulting in the linear growth of its cost with $\bar{\lambda}$ illustrated by the orange dashed line in Figure 5. The algorithms perform comparably up until around $\bar{\lambda} = 1$, after which the cost for our algorithm is always lower than the algorithm in [12]. Specifically, the observation/switching cost for the algorithm from [12] increases by 42.15 for every unit increase of $\bar{\lambda}$, since 42.15 is the expected number of switches under that algorithm. In contrast, in the regime explored in this simulation, the use of Algorithm 1 increases the observation/switching cost at a rate of around 16.3 per unit increase of $\bar{\lambda}$, meaning the cost of Algorithm 1 grows at a rate 61.3% slower than the [12] algorithm.

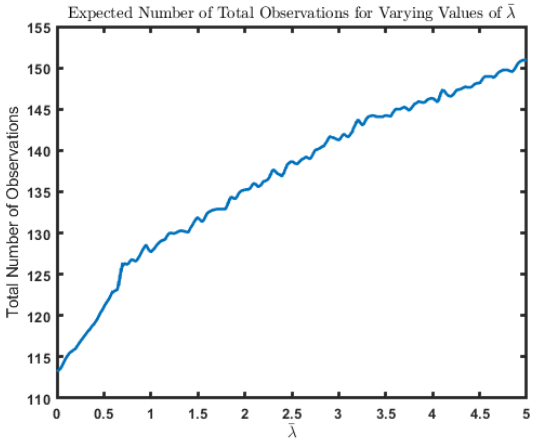

Fig. 4. The expected number of total observations before termination under Algorithm 1 for different values of $\bar{\lambda}$. The number of observations grows since a lower value of γ_L will require more observations be taken before Algorithm 1 switches away from a data stream. However, the reduction in switching cost outpaces the increase in observation cost.

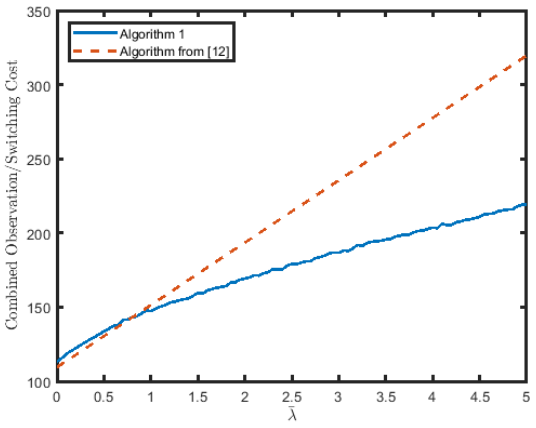

Fig. 5. The combined observation/switching costs for our Algorithm 1 (blue solid line) and the algorithm from [12] (orange dashed line).

VI. Conclusion

In this paper we introduced an algorithm which performs online anomaly search over many sequences with switching costs, with parameters that can be directly calculated, and almost uniform improvement over the best comparable method that does not account for switching costs [12]. We showed that the approximations used to derive this algorithm are accurate for problems of interest, and demonstrated the success of this algorithm with numerical simulations. Future work will embed the problem into a physical setting and perform optimal routing and control for a physical vehicle that incorporates the cost of switching that is due, e.g., to the downtime incurred by moving.

Acknowledgements

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any agency of the U.S. government. Examples of analysis performed within this article are only examples. Assumptions made within the analysis are also not reflective of the position of any U.S. Government entity. The Public Affairs approval number of this document is AFRL-2023-0996.

References

[1] H. Chernoff, "Sequential design of experiments," The Annals of Mathematical Statistics, vol. 30, no. 3, pp. 755–770, 1959.

[2] V. Dragalin, "A simple and effective scanning rule for a multi-channel system," Metrika, vol. 43, no. 1, pp. 165–182, 1996.

[3] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," IEEE Transactions on Automatic Control, vol. 58, no. 10, pp. 2451–2464, 2013.

[4] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," The Annals of Statistics, vol. 41, no. 6, pp. 2703–2738, 2013.

[5] S. Nitinawarat and V. V. Veeravalli, "Controlled sensing for sequential multihypothesis testing with controlled Markovian observations and non-uniform control cost," Sequential Analysis, vol. 34, no. 1, pp. 1–24, 2015.

[6] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," IEEE Transactions on Information Theory, vol. 61, no. 3, pp. 1432–1450, 2015.

[7] B. Huang, K. Cohen, and Q. Zhao, "Active anomaly detection in heterogeneous processes," IEEE Transactions on Information Theory, vol. 65, no. 4, pp. 2284–2301, 2018.

[8] A. Tsopelakos, G. Fellouris, and V. V. Veeravalli, "Sequential anomaly detection with observation control," in 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019, pp. 2389–2393.

[9] N. K. Vaidhiyan, S. P. Arun, and R. Sundaresan, "Neural dissimilarity indices that predict oddball detection in behaviour," IEEE Transactions on Information Theory, vol. 63, no. 8, pp. 4778–4796, 2017.

[10] N. K. Vaidhiyan and R. Sundaresan, "Active search with a cost for switching actions," in 2015 Information Theory and Applications Workshop (ITA). IEEE, 2015, pp. 17–24.

[11] T. Lambez and K. Cohen, "Anomaly search with multiple plays under delay and switching costs," IEEE Transactions on Signal Processing, vol. 70, pp. 174–189, 2021.

[12] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," IEEE Transactions on Information Theory, vol. 57, no. 8, pp. 5375–5386, 2011.

[13] F. Qin, H. Feng, T. Yang, and B. Hu, "Low-cost active anomaly detection with switching latency," Applied Sciences, vol. 11, no. 7, p. 2976, 2021.

[14] D. Chen, Q. Huang, H. Feng, Q. Zhao, and B. Hu, "Active anomaly detection with switching cost," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5346–5350.

[15] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an anomalous target with observations from an exponential family," in 2019 IEEE Data Science Workshop (DSW). IEEE, 2019, pp. 88–92.

[16] D. Siegmund, Sequential analysis: tests and confidence intervals. Springer Science & Business Media, 1985.