Sequential anomaly detection with observation control under a generalized error metric

Aristomenis Tsopelakos, Georgios Fellouris
Coordinated Science Lab, Department of Electrical and Computer Engineering, Department of Statistics
University of Illinois at Urbana-Champaign
Email: {tsopela2, fellouri}@illinois.edu

***Abstract*—The problem of sequential anomaly detection is considered under sampling constraints and generalized error control. It is assumed that there is no prior information on the number of anomalies. It is required to control the probability at least k errors, of any kind, upon stopping, where k is a user specified integer. It is possible to sample only a fixed number of processes at each sampling instance. The processes to be sampled are determined based on the already acquired observations. The goal is to find a procedure that consists of a stopping rule and a decision rule and a sampling rule that satisfy the sampling and error constraints, and have as small as possible average sample size for every possible scenario regarding the subset of anomalous processes. We characterize the optimal expected sample size for this problem to a first order approximation as the error probability vanishes to zero, and we propose procedures that achieve it. The performance of those procedures is compared in a simulation study for different values of k.**

Index Terms — Anomaly detection, generalized error, sampling design, asymptotic optimality.

I. Introduction

The need to identify a subset of anomalous or outlying processes arises in various contexts. For example, in economics the processes may refer to prices in stock market [9], while in fraud prevention security systems, they may refer to e-commerce activity [8]. In large scale systems, the practitioner may be willing to tolerate a small number of errors in the final decision in order to reach a conclusion faster. This tolerance to error can be expressed as a requirement to control the probabilities of at least k_1 false alarms and at least k_2 missed detections, or alternatively the probability of at least k errors, of any kind. We will refer to the former as control of generalized familywise error rates when either $k_1 > 1$ or $k_2 > 1$, and to the second as control of the generalized misclassification rate when $k > 1$.

Sequential procedures that control generalized familywise error rates were proposed in [11], [12], when all processes can be observed at each sampling instance. An asymptotic optimality theory for such procedures was developed in [5] for both generalized error metrics in [5].

When it is possible to sample only a fixed number of processes at each sampling instance, the sequential anomaly detection problem was considered with classical familywise error rates ($k_1 = k_2 = 1$) in [6]. In addition to a stopping rule that determines when to stop sampling, and a decision rule that determines which processes to declare as anomalous upon stopping, in our context it is also required to specify a sampling rule that determines which channels to observe at each sampling instance given the already collected observations. This formulation is also related to works such as [1], [2], [3], [13], [7].

In this paper, we combine these two lines of work and consider the sequential anomaly detection problem when it is possible to sample only a fixed number of processes at each sampling instance and it is required to control the generalized misclassification rate k without having any prior information on the size of the anomalous subset. We characterize the optimal expected sample size for this problem to a first-order asymptotic approximation as the error probability goes to 0, and we propose procedures that achieve it under every possible subset of anomalous processes.

To be more precise, we adopt the stopping and decision rule from [5] and we focus on the design of sampling rules that lead to asymptotic optimality. Specifically, we consider a probabilistic rule in the spirit of [3], [7], as well as a deterministic rule, which when we set $k = 1$, i.e, in the case of classical misclassification error, reduces to one of the sampling rules proposed in [6]. These rules are compared in a simulation study in which the deterministic rule exhibits better performance, especially when k is small, a result that agrees with previous findings in [1], [2], [6].

The remainder of the paper is organized as follows. In Section II we formulate the problem mathematically, in Section III we present the main result of this work, whereas in Section IV we introduce the proposed sampling rules. In Section V, we present the results of a simulation study which illustrates the performance of the proposed sampling rules for different values of k. In Section VI we discuss potential generalizations of this work.

II. Problem formulation

We consider M channels that generate observations sequentially in time, but at each time instance it is possible to sample only K of them, where K is a user-specified number in $[M] := \{1, \dots, M\}$. We must determine a sampling rule, that is a sequence of random vectors,

$$R(n) := (R_1(n), \dots, R_M(n)), \quad n \in \mathbb{N},$$

such that, for every $n \in \mathbb{N}$, $R(n)$ takes values in $\{0,1\}^M$, channel i is sampled at time n if and only if $R_i(n) = 1$,

$$R_1(n) + \ldots + R_M(n) = K, \tag{1}$$

and $R(n)$ is determined based on the available information from the first $n-1$ sampling instances. To be more precise, for each $n \in \mathbb{N}$ let $X_i(n)$ represent the observation from channel i at time n when $R_i(n) = 1$ and set $X(n) := (X_1(n), \ldots, X_M(n))$, where $X_i(n)$ is an arbitrary constant when $R_i(n) = 0$. Let also $Z(n)$ denote a random vector that is independent of the observations, and it is used for randomization purposes if necessary. Thus, $R(1)$ is arbitrary, but for every $n > 1$ the random variable $R(n)$ must be $\mathcal{F}_{n-1}$-measurable, where

$$\mathcal{F}_n := \sigma(X(s), Z(s), 1 \leq s \leq n). \tag{2}$$

For each $i \in [M]$ we assume that channel i is either anomalous or not, and in the former (resp. latter) case its observations have density g_i (resp. f_i) with respect to a measure ν_i. Specifically, if $A \subseteq [M]$ is the subset of anomalous channels, or simply the "anomalous subset", then

$$X_i(n) \,|\, R_i(n) = 1, \mathcal{F}_{n-1} \sim \begin{cases} g_i, & i \in A \\ f_i, & i \notin A. \end{cases}$$

We assume that the densities f_i, g_i, $i \in [M]$ are completely specified, and that the Kullback-Leibler (KL) information numbers,

$$I_i := \int \log(g_i/f_i)\, g_i \, d\nu_i, \quad D_i := \int \log(f_i/g_i)\, f_i \, d\nu_i, \tag{3}$$

are positive and finite. However, the anomalous subset is completely unknown, thus, there are 2^M distinct scenarios for it. To emphasize this, we use P_A and E_A to denote the underlying probability measure and the corresponding expectation when the anomalous subset is A.

Our goal is to stop sampling as quickly as possible and ideally identify, an anomalous subset upon stopping with less than k errors. We need to find a stopping rule that determines when to stop sampling, a decision rule that determines which channels to classify as anomalous upon stopping and a sampling rule, that determines the observations to be obtained from the channels. That is, we need to find a triplet (T, Δ, R) that consists of

- an $\mathbb{N}$-valued random variable, T, such that $\{T = n\} \in \mathcal{F}_n$ for every $n \in \mathbb{N}$,
- a sequence $\Delta := (\Delta(n), n \in \mathbb{N})$, where each $\Delta(n) := (\Delta_1(n), \ldots, \Delta_M(n))$ is an $\mathcal{F}_n$-measurable random vector that takes values in $\{0,1\}^M$,
- a sampling rule $R := (R(n), n \in \mathbb{N})$ which is $\mathcal{F}_{n-1}$-measurable.

so that sampling is terminated at time T and channel i is declared to be anomalous (resp. non-anomalous) upon stopping when $\Delta_i(T) = 1$ (resp. $\Delta_i(T) = 0$).

The problem we consider in this work is to find a triplet (T, Δ, R) such that $E_A[T]$ is as small as possible for every $A \subseteq [M]$, while guaranteeing that the probability of at least k mistakes, of any kind, is below α, where both $k \in [M-1]$ and $\alpha \in (0,1)$ are user-specified parameters. To be more specific, let $\mathcal{C}(\alpha; k)$ denote the family of triplets (T, Δ, R) for which

$$P_A(|\Delta(T) \triangle A| \geq k) \leq \alpha \quad \text{for every } A \subseteq [M]. \tag{4}$$

where $\triangle$ represents the symmetric difference of two sets, i.e., $\Delta(T) \triangle A := (A \setminus \Delta(T)) \cup (\Delta(T) \setminus A)$.

Let (T, Δ, R) be a procedure that belongs to $\mathcal{C}(\alpha; k)$ for any given α. We will say that such a procedure is asymptotically optimal under P_A if it achieves

$$\mathcal{J}_A(\alpha; k, K, M) := \inf_{(R,T,\Delta) \in \mathcal{C}(\alpha;k)} E_A[T] \tag{5}$$

to a first-order asymptotic approximation as $\alpha \to 0$, i.e., if

$$E_A[T] \sim \mathcal{J}_A(\alpha; k, K, M),$$

Since the anomalous subset is completely unknown, our goal in this work is to find a procedure that is asymptotically optimal for every $A \subseteq [M]$. This problem was considered in [5] when all channels are sampled at all times up to stopping $K = M$, in which case only a stopping and a decision rule need to be determined. Here, our focus is on the case of sampling constraints $K < M$ and our main goal is to obtain sampling rules that lead to asymptotic optimality when combined with the stopping and decision rule in [5].

III. Main result

A. Stopping and Decision rules

We start by describing the proposed stopping and decision rule given an arbitrary sampling rule $R(n), n \in \mathbb{N}$. For each $n \in \mathbb{N}$, we denote by $\Lambda_i(n)$ the log-likelihood ratio (LLR) of all observations in channel i up to some arbitrary time $n \in \mathbb{N}$, which takes the following form

$$\Lambda_i(n) = \sum_{s=1}^{n} \log\left(\frac{g_i(X_i(s))}{f_i(X_i(s))}\right) R_i(s). \tag{6}$$

We consider the absolute value of these LLRs

$$\bar{\Lambda}_i(n) := |\Lambda_i(n)|, \quad n \in \mathbb{N}, \tag{7}$$

and the corresponding order statistics,

$$\bar{\Lambda}_{(1)}(n) \leq \ldots \leq \bar{\Lambda}_{(M)}(n). \tag{8}$$

Following [5], we stop as soon as the sum of the k smallest absolute LLRs is larger than some positive threshold b, and we classify as abnormal any channel with non-negative log-likelihood ratio at the time of stopping. That is, the stopping time is

$$T_b^* := \inf\left\{ n \geq 1 : \sum_{i=1}^{k} \bar{\Lambda}_{(i)}(n) \geq b \right\} \tag{9}$$

and the proposed decision rule is

$$\Delta^*(n) := \{\hat{w}_1(n), \ldots, \hat{w}_{p(n)}(n)\}, \quad n \in \mathbb{N}, \tag{10}$$

where $p(n)$ is the number of non-negative LLRs at time n, and $\hat{w}_1(n), \dots, \hat{w}_{p(n)}(n)$ denote the indices of the increasingly ordered non-negative LLRs at time n, i.e.,

$$0 \leq \Lambda_{\hat{w}_1(n)}(n) \leq \dots \leq \Lambda_{\hat{w}_{p(n)}}(n) \tag{11}$$

When all sensors are sampled at all times $K = M$, in which case there is no need to specify a sampling rule, it follows from [5] that, $(T_b^*, \Delta^*) \in \mathcal{C}(\alpha; k)$ holds for any α when b is equal to or larger than

$$b(\alpha) = |\log(\alpha)| + \log \binom{M}{k}, \tag{12}$$

and $(T_{b(\alpha)}^*, \Delta^*)$ is asymptotically optimal for every $A \subseteq [M]$.

When we do not sample all sensors at all times $K < M$, the same arguments as in [5] can be used to show that $(T_{b(\alpha)}^*, \Delta^*, R)$ belongs to $\mathcal{C}(\alpha; k)$ for any sampling rule, R.

B. Asymptotic optimality

For the triplet $(T_{b(\alpha)}^*, \Delta^*, R)$ to be asymptotically optimal, we need to impose two conditions on the sampling rule, R.

Definition 1: A sampling rule R is consistent if for the random time

$$\sigma_A := \sup\{n \geq 1 : \Delta^*(n) \neq A\}. \tag{13}$$

it guarantees $E_A[\sigma_A] < \infty$, for any $A \subseteq [M]$.

The first condition is that the sampling rule must be consistent, i.e. to guarantee that the subset of anomalies is estimated correctly forever after a random time of finite mean value.

The second condition is that for every $i \in [M]$ we have

$$\sum_{n=1}^{\infty} P_A \left(\Big| \frac{1}{n} \sum_{s=1}^{n} R_i(s) - c_i(A) \Big| > \epsilon \right) < \infty, \ \forall \epsilon > 0 \tag{14}$$

where the quantities $c_1(A), \dots, c_M(A)$ have to satisfy the properties (18)-(20). The second condition determines which are the desired limiting sampling frequencies that the sampling rule must guarantee.

To state properties (18)-(20), let us denote by $F_i(A)$ the i^{th} smallest number in the set

$$\{I_i, D_j, i \in A, \, j \notin A\}, \tag{15}$$

and by $\widetilde{F}_i(A)$ the harmonic mean of the $M - i + 1$ largest numbers in the same set, i.e.,

$$\begin{aligned} &F_1(A) \leq \dots \leq F_M(A), \\ &\widetilde{F}_i(A) := \frac{M - i + 1}{\sum_{u=i}^{M} 1/F_u(A)}, \quad i \in [M]. \end{aligned} \tag{16}$$

For each $i \in [M]$ we also define the following quantity

$$Q_i(A) := (M - i + 1) \, \frac{F_i(A)}{\widetilde{F}_i(A)} + i - 1, \tag{17}$$

where $Q_0(A) := 0$ and we note that

$$Q_1(A) \leq \dots \leq Q_M(A) = M.$$

For every $i \in [M]$, let (i) denote the identity of the channel with the i^{th} smallest number in (15). It is more convenient to describe the desired limiting sampling frequency for channel (i) and not directly for channel i. Since we know the process to which $F_i(A)$ refers, once we have specified $c_{(i)}(A)$, $i \in [M]$, we can recover $c_i(A)$, $i \in [M]$ which are used in the sampling rules.

First, we require that

$$c_{(i)}(A) = 1, \quad \text{for} \quad 1 \leq i \leq m(A), \tag{18}$$

where $m(A)$ is defined as follows:

$$m(A) := \max\{0 \leq i \leq k : K \geq Q_i(A)\}$$

For the remaining limiting sampling frequencies we distinguish two cases depending on whether $m(A) < k$ or $m(A) = k$. Thus, for $m(A) < i \leq M$ we require that

$$c_{(i)}(A) = x(A) \, \frac{\widetilde{F}_{m(A)+1}(A)}{F_i(A)} \quad \text{when} \quad m(A) < k \tag{19}$$

$$c_{(i)}(A) \geq \frac{F_k(A)}{F_i(A)} \qquad \text{when} \quad m(A) = k, \tag{20}$$

where $x(A)$ is defined as follows

$$\frac{K - m(A)}{M - m(A)},$$

and is a non-negative quantity since by the definition of $m(A)$ we have

$$K \geq Q_{m(A)}(A) > m(A) - 1$$

We now state the main result of this work, which provides the first-order asymptotic approximation to the optimal expected sample size.

Theorem 1: Fix $k \in [M-1]$, $K \in [M]$ and $A \subseteq [M]$. Then, as $\alpha \to 0$

$$\mathcal{J}_A(\alpha; k, K, M)| \sim \frac{|\log \alpha|}{V_A(k, K, M)}, \tag{21}$$

where $V_A(k, K, M)$ is defined as follows

$$\sum_{u=1}^{m(A)} F_u(A) \, + \, (k - m(A)) \, x(A) \, \widetilde{F}_{m(A)+1}(A). \tag{22}$$

Moreover, if $b(\alpha)$ is given by (12), R is consistent and (14) holds with $c_1(A), \dots, c_M(A)$ implied by (18)-(20); then $(T_{b(\alpha)}^*, D^*, R)$ is asymptotically optimal under P_A.

Sketch of proof: Fix $A \subseteq [M]$. The first step is to show that

$$\mathcal{J}_A(\alpha; k, K, M) \geq \frac{|\log \alpha| \, (1 + o(1))}{\max_{\mathcal{D}_K} \mathcal{V}_A(c_1, \dots, c_M)},$$

where $o(1)$ is a term that goes to 0 as $\alpha \to 0$,

$$\mathcal{D}_K := \{(c_1, \dots, c_M) \in [0, 1]^M : c_1 + \dots + c_M = K\},$$

and $\mathcal{V}_A(c_1, \dots, c_M)$ is equal to

$$\min_{C \subseteq [M] : \, |A \triangle C| = k} \left(\sum_{i \in A \setminus C} c_i I_i + \sum_{j \in C \setminus A} c_j D_j \right).$$

which is equivalently expressed as

$$\min_{U \subseteq [M]:\, |U|=k} \sum_{i \in U} c_{(i)} F_A^i \tag{23}$$

The max-min structure of (23) implies that the maximum value of the function $\mathcal{V}_A$ on $\mathcal{D}_K$ is given by (22), and is achieved by $c_{(i)}$ that satisfy (18)-(20). The second step is to show that this asymptotic lower bound is achieved by a rule of the form $(T^*_{b(\alpha)}, \Delta^*, R)$, which is implied by the imposed conditions on the sampling rule. ∎

We proceed to implications of Theorem 1.

Corollary 1: If $m(A) = k$, equivalently $K \geq Q_k(A)$, then

$$V_A(k, K, M) = F_1(A) + \ldots + F_k(A),$$

and as a result for any $A \subseteq [M]$, as $\alpha \to 0$

$$\lim_{\alpha \to 0} \frac{\mathcal{J}_A(\alpha; k, M, M)}{\mathcal{J}_A(\alpha; k, K, M)} = 1.$$

Therefore, if $K < M$ but $K \geq Q_k(A)$, the first-order asymptotic approximation of the optimal expected sample size under P_A is the same as in the case of full sampling $K = M$.

The next corollary shows that the first-order asymptotic approximation of the optimal expected size is reduced by at least a factor of k compared to the case of no tolerance $k = 1$.

Corollary 2: For any $A \subseteq [M]$, as $\alpha \to 0$

$$\lim_{\alpha \to 0} \frac{\mathcal{J}_A(\alpha; k, K, M)}{\mathcal{J}_A(\alpha; 1, K, M)} \leq \frac{1}{k}.$$

Proof: It suffices to verify that for $k = 1$,

$$V_A(1, K, M) = cF_1(A)$$

where $c \leq 1$ is constant; while for $k > 1$,

$$V_A(k, K, M) \geq kcF_1(A)$$

∎

As a special case of the problem, we can consider

$$I_i = I \leq D = D_i, \quad i \in [M]. \tag{24}$$

Then, for every $A \subseteq [M]$ we have

$$Q_j(A) = M - (1 - I/D)\,(M - |A|) \quad \text{for} \quad j \leq |A|, \tag{25}$$
$$Q_j(A) = M, \quad \text{for} \quad j > |A. \tag{26}$$

and depending on K, we use (25) to calculate the c_i. When in particular $I = D$, for every $A \subseteq [M]$ we have $m(A) = 0$, $c_i = K/M$ and as a result

Corollary 3: Suppose $I = D$, then for every $A \subseteq [M]$ as $\alpha \to 0$

$$\mathcal{J}_A(\alpha; k, K, M) \sim \frac{|\log(\alpha)|}{k(K/M)I}$$

IV. Sampling rules

Theorem 1 suggests that a sampling rule should be consistent and designed so that for every $A \subseteq [M]$ it guarantees that (14) holds with limiting sampling frequencies, $c_i(A), i \in [M]$, implied by (18)-(20). In this section we propose three sampling rules, which satisfy these conditions.

To define the three sampling rules, we need to describe how each of them selects the K channels to be sampled at some arbitrary time $n + 1$ given the available information from the first n sampling instances. For all of them, this selection will rely on the estimate, $\Delta^*(n)$, of the anomalous subset after the first n sampling instances, defined in (9), and on the desired sampling frequencies, $c_i(A)$, $i \in [M]$ for every $A \subseteq [M]$ provided in (18)-(20).

When $m(A) = k$, or equivalently $K \geq Q_k(A)$, (20) does not determine uniquely $c_{(i)}(A) \in (0, 1)$ with $i > k$. A specific selection which aims to the maximization of the number of limiting sampling frequencies that are equal to 1, is provided by the following algorithm:

Step A: Initialize

$$c_{(i)}(A) = F_k(A)/F_i(A), \quad \forall\, i > k \tag{27}$$

Step B: Allocate the extra quantity $K - Q_k(A)$ to every $c_{(i)}(A) < 1$, $i \geq k + 1$, such that the equality

$$c_{(i)}(A)\; F_i(A) = c_{(j)}(A)\; F_j(A), \quad \forall\, i, j > k \tag{28}$$

is preserved. By the time a $c_{(i)}(A)$ becomes equal to 1, we do not consider it anymore in (28).

According to all three rules, we sample at $n + 1$ those channels whose desired limiting sampling frequencies are equal to 1 when the actual anomalous subset is the one estimated by $\Delta^*(n)$, i.e., all channels in

$$G_n := \{i \in [M] : c_i\,(\Delta^*(n)) = 1\}. \tag{29}$$

The three rules differ in how the remaining $K - |G_n|$ channels are sampled. In other words, they differ in how to select at time $n + 1$ a subset from

$$\mathcal{P}_n := \{B \in [M] \setminus G_n : |B| = K - |G_n|\}.$$

A. Chernoff rule

The first approach is to sample $B \in \mathcal{P}_n$ with probability $q_n(B)$, where q_n is a probability mass function (pmf) on $\mathcal{P}_n$ such that the probability that channel $i \in [M] \setminus G_n$ is selected at time $n + 1$ equals the desired limiting sampling frequency of this channel when the true anomalous subset is estimated as $\Delta^*(n)$, i.e.,

$$\sum_{B \in \mathcal{P}_n:\, i \in B} q_n(B) = c_i(\Delta^*(n)) \;\; \text{for every } i \in [M] \setminus G_n. \tag{30}$$

We observe that the larger the G_n the smaller the number of unknowns in (30), which reduces the computational complexity of (30).

We refer to this sampling rule as Chernoff rule, since in the case of no tolerance ($k = 1$) it is an improved version of the probabilistic sampling rule that is implied by the general framework in [3].

B. Bernoulli rule

A convenient modification of the Chernoff rule can be obtained if we sample at time $n+1$ each channel $i \in [M] \backslash G_n$ with probability $c_i(\Delta^*(n))$. We refer to this sampling rule as Bernoulli, since for each channel $i \in [M] \backslash G_n$ we need to draw a Bernoulli random variable $Z_i(n)$ with parameter $c_i(\Delta^*(n))$ and sample channel i at time $n+1$ if and only if $Z_i(n) = 1$. This sampling rule does not require the solution of the linear system (30), but it does not guarantee that exactly K channels are sampled at each time, therefore it does not satisfy constraint (1). Nevertheless, the average number of observations per sampling instance converges to K exponentially fast when sampling continues indefinitely, as we can see for example by Hoeffding's inequality. Therefore, this rule may be acceptable in practice when it is not crucial to respect the constraint (1) at all times, but it suffices to guarantee that K observations are taken on average per sampling instance.

C. Equalizer rule

Contrary to the two previous sampling rules where the selection of a subset from $\mathcal{P}_n$ is the result of a random mechanism, the third sampling rule selects all channels at $n+1$, apart from at most one, in a deterministic way given the observations up to time n. Specifically, let N_n denote the sum of the desired limiting sampling frequencies from the channels in $\Delta^*(n) \setminus G_n$, i.e.

$$N_n := \sum_{i \in \Delta^*(n) \setminus G_n} c_i(\Delta^*(n)). \tag{31}$$

The Equalizer sampling rule samples at time $n+1$

- All processes in G_n.
- For the remaining processes in $[M] \setminus G_n$, we sample the $\lfloor N_n \rfloor$ processes with the smallest non-negative LLRs and the $K - |G_n| - \lceil N_n \rceil$ processes with the largest negative LLRs.
- If N_n is not an integer, we draw an independent Bernoulli random variable $Z(n)$ with parameter $N_n - \lfloor N_n \rfloor$. Among the processes not sampled in the first two items we sample the one with the smallest non-negative (resp. largest negative) LLR if $Z(n) = 1$ (resp. 0).

We refer to this rule as the Equalizer rule, since it forces all the LLR $\Lambda_i(n)$, $i \in [M] \setminus G_n$ to stay close together.

V. Simulation Study

In this section we present a simulation study with $M = 10$, $K = 5$, $f_i = \mathcal{N}(0, 1)$ and $g_i = \mathcal{N}(\mu_i, 1)$, thus, $I_i = D_i = (\mu_i)^2/2$, where $i \in [M]$. We set

$$\mu_i = \begin{cases} 0.5, & 1 \leq i \leq 3 \\ 0.7, & 4 \leq i \leq 7 \\ 1, & 8 \leq i \leq 10 \end{cases}$$

We compare the performance of the three sampling rules that we introduced in the previous section for different values of $k \in [M-1]$ when A contains the five first processes and $\alpha = 10^{-3}$. Moreover, in our simulation study we did not select b according to (12), but we performed simulation experiments to find those threshold values for each sampling rule for which the error constraint is satisfied with approximate equality.

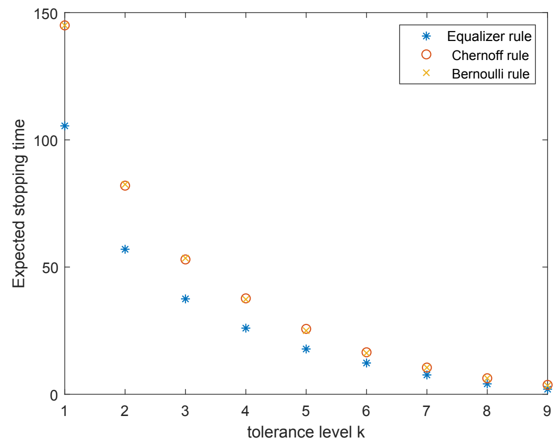

Fig. 1. Expected stopping time versus the tolerance level k.

In Fig. 1, we plot the estimated expected sample size that corresponds to each sampling rule against k, with maximum standard error 10^{-2}. We see that the performance of the Equalizer rule is better, especially when k is small, whereas the Chernoff rule and the Bernoulli rule exhibit similar performance. For any rule, the expected stopping time is reduced by a factor of at least $1/k$ as k increases; a fact implied by Corollary 2.

VI. Conclusion

We consider the sequential anomaly detection problem under sampling constraints, where we can observe only a predetermined number of channels at each instance. We assume a generalized error metric where we tolerate up to $k-1$ misclassification errors. We characterize asymptotically optimal sampling rules and we compute the asymptotic optimal performance. We provide three sampling rules that achieve the asymptotically optimal performance, but differ significantly with respect to their computational complexity and performance in finite regime for all possible values of $k < M$. We perform a simulation which depicts the better performance of the Equalizer rule compared to the Chernoff and the Bernoulli rule, in finite regime.

In this work we have assumed for simplicity that the anomalous and non-anomalous behavior in each process is completely specified. A first direction of generalization of the current work would be to remove this assumption. Moreover, it is interesting to consider the same problem in the case of generalized familywise error rates.

VII. Acknowledgements

This work was supported in part by National Science Foundation through grants CIF 1514245 and DMS ATD 1737962.

REFERENCES

[1] Kobi Cohen, and Zhao Qing. "Active hypothesis testing for anomaly detection." IEEE Transactions on Information Theory 61.3 (2015): 1432-1450.
[2] Boshuang Huang, Kobi Cohen, and Zhao Qing. "Active anomaly detection in heterogeneous processes." IEEE Transactions on Information Theory (2018).
[3] Herman Chernoff, "Sequential design of experiments." The Annals of Mathematical Statistics 30.3 (1959): 755-770.
[4] Yanglei Song, and Georgios Fellouris, "Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals." Electronic Journal of Statistics 11.1 (2017): 338-363.
[5] Yanglei Song and Georgios Fellouris. "Sequential multiple testing with generalized error control: An asymptotic optimality theory." The Annals of Statistics 47.3 (2019): 1776-1803.
[6] Aristomenis Tsopelakos, Georgios Fellouris, and Venugopal V. Veeravalli. "Sequential anomaly detection with observation control." 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019.
[7] Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. "Controlled sensing for multihypothesis testing." IEEE Transactions on Automatic Control 58.10 (2013): 2451-2464.
[8] Richard J. Bolton, and David J. Hand. "Statistical fraud detection: A review." Statistical science (2002): 235-249.
[9] Elroy Dimson, ed. Stock market anomalies. CUP Archive, 1988.
[10] Abraham Wald, Sequential analysis. Courier Corporation, 1973.
[11] Shyamal K. De, and Michael Baron, "Sequential tests controlling generalized familywise error rates." Statistical Methodology 23 (2015): 88-102.
[12] Jay Bartroff, "Multiple hypothesis tests controlling generalized error rates for sequential data." Statistica Sinica 28.1 (2018): 363-398.
[13] S. A. Bessler, Theory and Applications of the Sequential Design of Experiments, Actions and Infinitely Many Experiments: Part 1-Theory, Dept. of Statistics, Stanford Univ., Tech. Report no. 55, 1960.