Sequential anomaly detection with observation control

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

***Abstract*—The problem of anomaly detection is considered when multiple processes are observed sequentially, but it is possible to sample only a subset of them at a time according to an adaptive sampling policy. The problem is to stop sampling as soon as possible and identify the anomalous processes, while controlling appropriate error probabilities. We consider two versions of this problem: in the first one there is no assumption regarding the anomalous processes, in the second their number is assumed to be known a priori. For each version, we obtain the optimal asymptotic performance as the error probabilities vanish and characterize the sampling rules that lead to asymptotic optimality. Moreover, we present two sampling rules for each setup, which differ in terms of the computational complexity and the actual performance they imply.**

Index Terms — Anomaly detection, outlying sequence detection, sequential design of experiments, asymptotic optimality.

I. Introduction

The need to identify a subset of anomalous or outlying processes arises in various contexts. For example, in biology the processes may refer to brain cell signals [9], while in economics the processes may refer to prices in stock market [10]. In many of these applications, the observations are obtained in a sequential manner and there are constraints on the number of processes that can be observed at each time step. The question that arises is how to accurately determine the subset of anomalous processes as quickly as possible subject to such sampling constraints.

When all processes are observed at each time, this problem has been studied in [6] when an upper and a lower bound on the number of anomalies are available. When there are sampling constraints and the number of anomalies is known a priori, this problem has been considered in [1], [2]. Specifically, the special case of homogeneous processes is considered in [1], whereas the general, non-homogeneous case is considered in [2] when either only *one* process can be observed at a time, or there is exactly *one* anomalous process. The related problem of testing multiple hypotheses with observation control was considered in [4], where the groundbreaking work of Chernoff [3] was extended beyond the case of two hypotheses. A dynamic programming approach to this problem was presented in [7]. For other detection problems with observation control we refer to [8].

In this paper, we consider the above anomaly detection problem when the number of anomalies is known a priori, but also in the more robust setup where the anomalous subset is completely unknown. For each of these two setups, we obtain the optimal asymptotic performance as the error probabilities vanish and characterize the sampling rules that lead to asymptotic optimality. Moreover, we present two different sampling rules for each setup. The first one is based on [4] and samples each possible subset of processes with some probability. The second requires ordering the local test statistics that correspond to the various processes and samples those processes with the smallest positive or largest negative values. When the number of anomalies is known, the latter approach generalizes the sampling rules in [1], [2], whereas to the best of our knowledge it is novel in the case that the anomalous subset is completely unknown.

The rest of the paper is organized as follows. In Section II, we formulate the problem mathematically. In Section III, we introduce the proposed stopping and decision rules. In Section IV, we obtain the optimal asymptotic performance and characterize the sampling rules that achieve asymptotic optimality. In Section V, we present two examples of such sampling rules. We conclude in Section VI. The proofs are not presented due to lack of space.

II. Problem formulation

We assume there are M channels, each of which is either anomalous or not. An observation from channel i is modeled as a random variable with density g_i when the channel is anomalous and density f_i otherwise, where $i \in [M] \equiv \{1, \ldots, M\}$. Here, f_i, g_i are densities with respect to a σ-finite measure ν_i for which the corresponding Kullback-Leibler information numbers

$$I_i = \int \log(g_i/f_i)\, g_i \, d\nu_i, \quad D_i = \int \log(f_i/g_i)\, f_i \, d\nu_i \quad (1)$$

are positive and finite for every $i \in [M]$. The problem is to identify the anomalous channels as quickly as possible when only K of the M channels can be sampled at a time. That is, if $R_i(n)$ is equal to 1 if channel i is sampled at time $n \in \mathbb{N} \equiv \{1, 2, \ldots\}$ and 0 otherwise, we must have

$$R_1(n) + \ldots + R_M(n) = K. \quad (2)$$

We denote by $X_i(n)$ the observation from channel i at time n when $R_i(n) = 1$, whereas $X_i(n)$ is defined in an arbitrary

way when $R_i(n) = 0$. We set $R(n) = (R_1(n), \ldots, R_M(n))$, $X(n) = (X_1(n), \ldots, X_M(n))$, and also denote by $Z(n)$ a random vector that is independent of the previous observations and may be used for randomization purposes. We require that $R(n+1)$ be $\mathcal{F}_n$-measurable for every $n \geq 1$, where $\mathcal{F}_n$ is the σ-algebra that contains all available information up to time n, i.e.,

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

Then, if A is the subset of anomalous channels, we have:

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

In addition to a sampling rule, we need to determine a stopping rule T and a decision rule Δ, where T is an $\{\mathcal{F}_n\}$-stopping time at which we stop sampling, i.e., $\{T = n\} \in \mathcal{F}_n$ for every $n \geq 1$, and $\Delta = (\Delta_1, \ldots, \Delta_M)$ is an $\mathcal{F}_T$-measurable random vector such that Δ_i is 1 if channel i is declared to be anomalous upon stopping and 0 otherwise. We want to select a triplet (R, Δ, T) so that the expected value of T is small and the true subset of anomalous channels is identified upon stopping. We will consider two formulations for this problem. Here, and in what follows, we denote by P_A and E_A the underlying probability measure and the corresponding expectation when the subset of anomalous channels is $A \subseteq [M]$.

A. Two versions of the problem

In the first version of the problem, the subset of anomalous channels is completely unknown. We denote by $C_{\alpha,\beta}$ the class of triplets (R, T, D) for which the probability of at least one *false alarm* and at least one *missed detection* is below α and β respectively, i.e.,

$$P_A\left(\bigcup_{i \notin A} \{\Delta_i = 1\}\right) \leq \alpha \quad \& \quad P_A\left(\bigcup_{i \in A} \{\Delta_i = 0\}\right) \leq \beta$$

for every $A \subseteq [M]$, where α, β are user-specified probabilities. The problem then is to find a procedure in $C_{\alpha,\beta}$ that achieves

$$\inf_{(R,T,\Delta) \in C_{\alpha,\beta}} E_A[T] \tag{3}$$

to a first-order asymptotic approximation as $\alpha, \beta \to 0$ for every $A \subseteq [M]$. For simplicity, we further assume that there is a positive real number $r > 0$ such that

$$|\log \alpha| \sim r \,|\log \beta|$$

as $\alpha, \beta \to 0$. Thus, r is a parameter that expresses the asymmetry in the user-specified tolerance levels.

In the second version of the problem, the number of anomalies is known *a priori* to be equal to L, and consequently the final decision rule will specify L channels as anomalous. In this case, if there is a false identification, there is both a false alarm and a missed detection and it is more relevant to control the probability of *misclassification*. Specifically, we denote by C_γ the class of procedures (R, T, Δ) for which the probability of at least one error of any kind is at most γ, i.e.,

$$P_A\left(\bigcup_{i \in A, j \notin A} \{\Delta_i = 0\} \cup \{\Delta_j = 1\}\right) \leq \gamma$$

for every $A \subseteq [M]$ with size $|A| = L$, where again $\gamma \in (0, 1)$ is a user-specified target level. Then, the second problem we consider is to find a procedure in C_γ that achieves

$$\inf_{(T,\Delta,R) \in C_\gamma} E_A[T] \tag{4}$$

to a first-order asymptotic approximation as $\gamma \to 0$ for every $A \subseteq [M]$ such that $|A| = L$.

B. Notation

For each $A \subseteq [M]$ we set

$$\begin{aligned} I_A^* &= \min_{i \in A} I_i, & D_A^* &= \min_{i \notin A} D_i \\ I_A &= \frac{|A|}{\sum_{i \in A}(1/I_i)}, & D_A &= \frac{M - |A|}{\sum_{i \notin A}(1/D_i)}. \end{aligned} \tag{5}$$

That is, I_A^* (resp. D_A^*) is the minimum and I_A (resp. D_A) the *harmonic* mean of the numbers in $\{I_i, i \in A\}$ (resp. $\{D_i, i \notin A\}$).

We denote by $\Lambda_i(n)$ the log-likelihood ratio (LLR) of all observations in channel i up to time n:

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

We also consider the corresponding order statistics

$$\Lambda_{(1)}(n) \geq \Lambda_{(2)}(n) \geq \ldots \geq \Lambda_{(M)}(n) \tag{7}$$

and denote by $w_1(n), \ldots, w_M(n)$ the corresponding indices:

$$\Lambda_{(i)}(n) = \Lambda_{w_i(n)}, \quad i \in [M]. \tag{8}$$

We denote by $p(n)$ the number of non-negative LLRs at time n and by $\hat{w}_1(n), \ldots, \hat{w}_{p(n)}(n)$ the indices of the increasingly ordered positive LLRs at time n.

$$0 \leq \Lambda_{\hat{w}_1(n)}(n) \leq \ldots \leq \Lambda_{\hat{w}_{p(n)}}(n). \tag{9}$$

For each $i \in [M]$ we denote by $\pi_i(n)$ the proportion of times channel i has been sampled up to time n, i.e.,

$$\pi_i(n) = \frac{1}{n} \sum_{s=1}^{n} R_i(s), \quad n \geq 1.$$

We say that $(z_1, \ldots, z_M)$ is a vector of *limiting sampling frequencies* under P_A if for every $\epsilon > 0$ and $i \in [M]$ we have

$$\sum_{n=1}^{\infty} P_A\left(|\pi_i(n) - z_i| > \epsilon\right) < \infty, \tag{10}$$

i.e., if z_i is the long-run proportion of times channel i is sampled under P_A when sampling is continued indefinitely. Note that due to the sampling constraint (2), a vector of limiting sampling frequencies must belong to

$$\mathcal{D}_K = \{(c_1, \ldots, c_M) \in [0, 1]^M : c_1 + \ldots + c_M = K\}. \tag{11}$$

III. Main result

A. Stopping and decision rules

When the number of anomalies is unknown, we follow the Intersection rule proposed in [11] and stop as soon as every LLR is outside $(-a, b)$, where a, b are positive thresholds to be determined, and claim that the anomalous channels are the ones with positive LLRs upon stopping. In other words, the stopping rule is

$$T_{int} = \inf\{n \geq 1 : \Lambda_i(n) \notin (-a, b) \; \forall \, i \in [M]\} \tag{12}$$

and the decision rule

$$\Delta_{int} = \{\hat{w}_1(T_{int}), \ldots, \hat{w}_{p(T_{int})}(T_{int})\}. \tag{13}$$

It is known [6] that when all M sensors are sampled at each time ($K = M$) and the thresholds are

$$a = |\log(\beta)| + \log(M), \quad b = |\log(\alpha)| + \log(M), \tag{14}$$

the resulting procedure belongs to $C_{\alpha,\beta}$ and achieves (3) as $\alpha, \beta \to 0$; in particular, for every $A \subseteq [M]$ we have

$$E_A[T_{int}] \sim \frac{|\log(\alpha)|}{J_A^*} \sim \inf_{(T,\Delta) \in C_{\alpha,\beta}} E_A[T], \tag{15}$$

where

$$J_A^* = \begin{cases} rD_A^*, & |A| = 0, \\ \min\{I_A^*, rD_A^*\}, & 0 < |A| < M, \\ I_A^*, & |A| = M. \end{cases} \tag{16}$$

On the other hand, when the number of anomalies is known to be L, where $0 < L < M$, we stop as soon as the "gap" between the L^{th} and the $(L+1)^{th}$ largest LLRs exceeds some positive value d and we claim that the anomalous channels are the L ones with the largest LLRs. In other words, the stopping rule is

$$T_{gap} = \inf\{n \geq 1 : \Lambda_{(L)}(n) - \Lambda_{(L+1)}(n) \geq d\} \tag{17}$$

and the decision rule is

$$\Delta_{gap} = \{w_1(T_{gap}), \ldots, w_L(T_{gap})\}. \tag{18}$$

It is known [6] that when all sensors are sampled at all times ($K = M$) and threshold d is selected so that

$$d = |\log(\gamma)| + \log(L(M - L)), \tag{19}$$

the resulting procedure belongs to C_γ and achieves (4) as $\gamma \to 0$; in particular, for every $A \subseteq [M]$ with $|A| = L$ we have

$$E_A[T_{gap}] \sim \frac{|\log(\gamma)|}{I_A^* + D_A^*} \sim \inf_{(T,\Delta) \in C_\gamma} E_A[T]. \tag{20}$$

The main goal of the present paper is to extend the above results to the case that we sample only K channels at each time, where $1 \leq K < M$.

B. Unknown number of anomalies

For each $A \subseteq [M]$, we introduce the function

$$J_A(c_1, \ldots, c_M) = \min\left\{\min_{i \in A}(c_i I_i), \; r \min_{j \notin A}(c_j D_j)\right\}, \tag{21}$$

with $(c_1, \ldots, c_M) \in \mathcal{D}_K$, defined in (11). We denote by $(c_1^A, \ldots, c_M^A)$ the optimizer of J_A, which satisfies:

$$I_i \, c_i^A = rD_j \, c_j^A = \min\{K/K_A, 1\} \; J_A^* \tag{22}$$

for every $i \in A$ and $j \notin A$, where J_A^* is defined in (16) and K_A is equal to

$$(M - |A|) \; \frac{\min\{I_A^*, rD_A^*\}}{rD_A} + |A| \; \frac{\min\{I_A^*, rD_A^*\}}{I_A} \tag{23}$$

when $0 < |A| < M$ and equal to

$$\begin{array}{ll} M \; I_A^*/I_A, & \text{when} \quad |A| = M \\ M \; D_A^*/D_A, & \text{when} \quad |A| = 0. \end{array} \tag{24}$$

The following theorem says that any sampling policy whose limiting sampling frequencies satisfy (22) leads to asymptotic optimality.

Theorem 1: Consider a procedure that consists of the stopping rule (12) with thresholds (14), the decision rule (13), and a sampling rule with limiting sampling frequencies given by (22). Then, as $\alpha, \beta \to 0$ we have

$$E_A[T_{int}] \sim \frac{|\log(\alpha)|}{\min\{K/K_A, 1\} \; J_A^*} \sim \inf_{(R,T,\Delta) \in C_{\alpha,\beta}} E_A[T].$$

Comparing with (15), we conclude that the best first-order asymptotic performance increases by a factor of $\max\{K_A/K, 1\}$ when sampling K, instead of M, channels at a time. By its definition in (23)–(24) it is clear that $K_A \leq M$ for every $A \subseteq [M]$ and that the equality holds when the problem is *homogeneous* and *symmetric*, in the sense that

$$I_i = I \quad \& \quad D_i = D, \quad \text{for every } i \in [M], \tag{25}$$

$$r = 1 \quad \& \quad D_i = I_i \quad \text{for every } i \in [M]. \tag{26}$$

Indeed, when (25) and (26) hold, $K_A = M$ for every $A \subseteq [M]$ and the limiting sampling frequencies in (22) become K/M, which means that all channels need to be sampled equally in the long run. On the other hand, if either (25) or (26) is violated, then K_A may be smaller than M, and in this case the ceiling of K_A is the minimum number of channels that need to be sampled at each time in order to achieve the best asymptotic performance when the subset of anomalous channels is A.

C. Known number of anomalies

When the number of anomalies is known to be L, we define the function

$$Q_A(c_1, \ldots, c_M) = \min\left\{\min_{i \in A}(c_i I_i) + \min_{i \notin A}(c_i D_i)\right\}, \tag{27}$$

where $(c_1, \ldots, c_M) \in \mathcal{D}_K$, defined in (11). As before, we denote by $(c_1^A, \ldots, c_M^A)$ the optimizer of Q_A, which is now given by the following identities:

$$c_i^A \, I_i = x_A \, I_A^* \qquad \text{and} \qquad c_j^A \, D_j = y_A \, D_A^* \tag{28}$$

for every $i \in A$ and $j \notin A$, where

$$x_A = \begin{cases} \min\{K/\hat{K}_A, \, 1\}, & \text{if } (M-L)I_A \geq LD_A \\ \min\{(K-\check{K}_A)^+/\hat{K}_A, 1\}, & \text{otherwise} \end{cases}$$
$$y_A = \begin{cases} \min\{(K-\hat{K}_A)^+/\check{K}_A, 1\}, & \text{if } (M-L)I_A \geq LD_A \\ \min\{K/\check{K}_A, \, 1\}, & \text{otherwise} \end{cases} \tag{29}$$

and

$$\hat{K}_A = L\frac{I_A^*}{I_A}, \qquad \check{K}_A = (M-L)\frac{D_A^*}{D_A}. \tag{30}$$

Theorem 2: Consider a policy that consists of the stopping rule (17) with threshold (19), the decision rule (18), and a sampling rule with limiting sampling frequencies given by (28). Then, as $\gamma \to 0$ we have

$$E_A[T_{gap}] \sim \frac{|\log(\gamma)|}{x_A I_A^* + y_A D_A^*} \sim \inf_{(R,T,\Delta)\in C_\gamma} E_A[T],$$

where x_A, y_A are defined in (29)-(30).

Similarly to the previous setup, the ceiling of $\hat{K}_A + \check{K}_A$ can be interpreted as the smallest number of channels that needs to be sampled at each time in order to achieve the best asymptotic expected sample size when the subset of anomalous channels is A. By (30) it is clear that $\hat{K}_A \leq L$ and $\check{K}_A \leq M - L$ for every $A \subseteq [M]$ and that the equalities hold in the homogeneous case where (25) holds.

IV. Sampling rules

In order to describe a sampling rule, we need to explain how at each time n we select the channels to be sampled at time $n+1$. The first step for this determination is the estimation of the anomalous channels at time n. When the number of anomalies is unknown, the proposed estimate is the set of *channels with positive LLR at time* n, i.e., $\{\hat{w}_1(n), \ldots, \hat{w}_{p(n)}(n)\}$, whereas when the number of anomalies is known to be L, the proposed estimate is the set of channels with the largest L LLRs at time n, i.e., $\{w_1(n), \ldots, w_L(n)\}$. Clearly, in each case the estimate *may vary with n and may not agree with the true subset of anomalies.* However, in order to lighten the notation, in what follows we simply denote it by A. We will present two sampling rules that guarantee the desired limiting sampling frequencies in different ways.

A. Probabilistic sampling

The first sampling rule is inspired by the approach in [3], [4] and samples a subset of channels of size K, say B, with some probability $q_A(B)$, where q_A is a pmf on the class of all subsets of $[M]$ with size K, which we will denote by $\mathcal{P}_K$. Specifically, the pmf q_A should be a solution of the following linear system with M equations and $\binom{M}{K}$ unknowns:

$$\sum_{B\in\mathcal{P}_K : i\in B} q_A(B) = c_i^A, \quad i\in[M], \tag{31}$$

where the c_i^As are given by (22) (resp. (28)) when the number of anomalies is unknown (resp. known). Note that the right-hand side in (31) is the probability that channel i is included in the selected sample when sampling according to q_A.

The computational complexity of solving this system is at least *cubic* in $\binom{M}{K}$. One may solve *offline* the system in (31) for every possible subset A and store the results for online implementation. Of course, this task becomes increasingly intensive in computation and demanding in memory as M increases, especially if the number of anomalies is unknown.

There are certain special cases in which the solution of this system is unique and straightforward. Consider for example the case where the number of anomalies is unknown: when (25) and (26) hold, then q_A is the uniform pmf for every A; when $M = 3$ and $K = 2$ subset $\{i, j\}$ is sampled with probability $(c_i^A + c_j^A - c_k^A)/2$, where $\{i, j, k\}$ is any permutation of $\{1, 2, 3\}$.

B. Deterministic sampling

The second sampling rule generalizes the approach proposed in [1] and samples at each time the channels with the smallest positive and/or smallest negative LLRs. Specifically, let

$$N_A = \sum_{i\in A} c_i^A, \tag{32}$$

where c_i^A is given by (22) (resp. (28)) when the number of anomalies is unknown (resp. known). If N_A is an integer, then we sample at $n+1$ the N_A channels with the smallest positive LLRs at time n and the $K - N_A$ channels with the largest negative LLRs at time n.

In the general case that N_A may not be an integer, we let $Z(n)$ be a Bernoulli random variable with parameter $N_A - \lfloor N_A \rfloor$. Then, if $Z(n) = 0$ (resp. $Z(n) = 1$), we sample at time $n+1$ the $\lfloor N_A \rfloor$ (resp. $\lceil N_A \rceil$) channels with the smallest positive LLRs at time n and the $K - \lfloor N_A \rfloor$ (resp. $K - \lceil N_A \rceil$) channels with the largest negative LLRs at time n.

This sampling rule determines how to sample at most $\lceil K_A \rceil$ (resp. $\lceil \hat{K}_A + \check{K}_A \rceil$) channels when the number of anomalies is unknown (resp. known). Any remaining sensors can be chosen arbitrarily. Most importantly, this sampling rule requires only the ordering of the LLRs, therefore it is easy to implement for any values of K and M, as its computational complexity at any given time is $O(M \log(M))$.

C. An illustration

In order to illustrate these sampling rules we consider a simple setup with $M = 3$ and $K = 2$, $f = N(0, 1)$, $g = N(0.5, 1)$, $r = 1$. Thus, $D_j = I_i = 1/8$ for every $i \in [M]$, which means that we have a homogeneous and symmetric setup, i.e., conditions (25) and (26) are satisfied.

When the number of anomalies is unknown, the probabilistic approach samples *at any time* uniformly over $\{1,3\},\{2,3\},\{1,2\}$. On the other hand, the sampling of the deterministic approach depends on the size of the current estimate of the subset of anomalous channels. Indeed, let A denote the estimated subset of anomalies at time n, i.e., the subset of channels with positive LLRs at time n. Then, $N_A = K|A|/M = 2|A|/3$, which implies that the 2 channels with the largest negative (resp. smallest positive) LLRs are sampled if $|A| = 0$ (resp. $|A| = 3$). When $|A| = 1$ (resp. $|A| = 2$) the largest negative (resp. smallest positive) is sampled always while the smallest positive (resp. largest negative) with probability $2/3$ and the second largest negative (resp. second smallest positive) with probability $1/3$.

When the number of anomalies is known to be $L \in \{1,2\}$, the condition $(M-L)I_A \geq LD_A$ reduces to $M \geq 2L$ for every subset of size L, thus, both sampling rules are independent of the estimated subset of anomalies in this case. Specifically, the deterministic rule samples the 2 channels with the largest (resp. smallest) LLRs when $L = 1$ (resp. $L = 2$), whereas the probabilistic approach is equally likely to sample each of the two subsets than contain the channel with largest (resp. smallest) LLR when $L = 1$ (resp. $L = 2$).

D. Simulation

We now present a simulation study with $M = 10$, $K = 5$, $f = N(0,1)$, $g = N(0.5,1)$, where we compare the Intersection rule (with $a = b$), which does not make any assumption regarding the anomalous subset, and the Gap rule, which assumes that the number of anomalies is known. Moreover, we implement each of these procedures with both the probabilistic and the deterministic sampling rule. In Fig. 1 we plot for each of the resulting four schemes the required expected sample size (ESS) versus the true number of anomalies when the probability of a false alarm/missed detection of the Intersection rule and the probability of misclassification of the Gap rule are approximately equal to 10^{-3}.

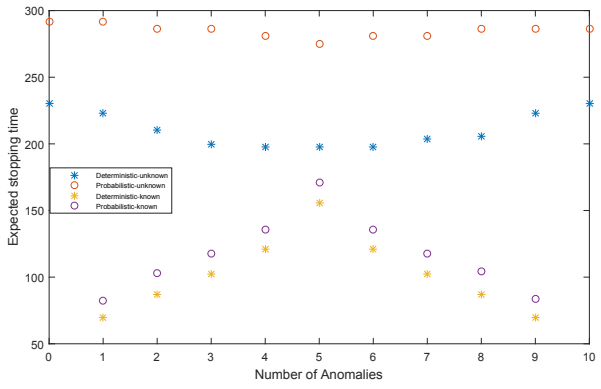

Fig. 1. Expected stopping time versus the number anomalies.

The first conclusion that we can draw from this figure is that we always obtain better performance using the deterministic instead of the probabilistic sampling rule. The difference is relatively small when the number of anomalies is known, but it is much more pronounced when the subset of anomalies is completely unknown. Another interesting observation is that the ESS of the Gap rule increases significantly (with both sampling rules) as the number of anomalies approaches $M/2 = 5$. On the other hand, the ESS of the Intersection rule is relatively stable as the number of anomalies varies.

V. Discussion

The above simulation example suggests that even if both sampling rules lead to *first-order* asymptotic optimality, the deterministic sampling rule leads to better performance in practice than the probabilistic one, at least in a homogeneous and symmetric setup where (25)-(26) hold. In view of the fact that it is also much easier to implement for large values of M and in non-homogeneous cases, we can argue that the deterministic sampling rule should be the default sampling rule for this problem.

Another important point is that a known number of anomalies is a quite restrictive and often unrealistic assumption that may lead to non-valid results in some cases of misspecification (e.g., if it is assumed that there are two anomalies but in reality there is only one). This paper shows that the anomaly detection problem admits an elegant, simple and efficient solution even if nothing is assumed about the anomalous subset. In current and future work, we plan to extend these results in the spirit of [6] and incorporate more realistic cases of prior information, such as an upper bound on the size of the anomalous subset.

VI. Acknowledgments

This work was supported by the National Science Foundation (NSF) under grants CIF 1514245 and DMS 1737962.

References

[1] Cohen, Kobi, and Qing Zhao. "Active hypothesis testing for anomaly detection." IEEE Transactions on Information Theory 61.3 (2015): 1432-1450.
[2] Huang Boshuang, Cohen Kobi, and Qing Zhao. "Active anomaly detection in heterogeneous processes." IEEE Transactions on Information Theory (2018).
[3] Chernoff, Herman. "Sequential design of experiments." The Annals of Mathematical Statistics 30.3 (1959): 755-770.
[4] Nitinawarat, Sirin, George K. Atia, and Venugopal V. Veeravalli. "Controlled sensing for multihypothesis testing." IEEE Transactions on Automatic Control 58.10 (2013): 2451-2464.
[5] Li, Yun, Sirin Nitinawarat, and Venugopal V. Veeravalli. "Universal outlier hypothesis testing." IEEE Transactions on Information Theory 60.7 (2014): 4066-4082.
[6] Song, Yanglei, and Fellouris Georgios, "Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals." Electronic Journal of Statistics 11.1 (2017): 338-363.
[7] Naghshvar, Mohammad, and Tara Javidi. "Active sequential hypothesis testing." The Annals of Statistics 41.6 (2013): 2703-2738.
[8] Tajer, Ali, and H. Vincent Poor. "Quick search for rare events." IEEE Transactions on Information Theory 59.7 (2013): 4462-4481.
[9] Stiles, Joan, and Terry L. Jernigan. "The basics of brain development." Neuropsychology review 20.4 (2010): 327-348.
[10] Dimson, Elroy, ed. Stock market anomalies. CUP Archive, 1988.
[11] De, Shyamal K., and Michael Baron. "Sequential Bonferroni methods for multiple hypothesis testing with strong control of family-wise error rates I and II." Sequential Analysis 31.2 (2012): 238-262.