Searching for Unknown Anomalies in Hierarchical Data Streams

Tomer Gafni, Kobi Cohen, *Senior Member, IEEE*, and Qing Zhao, *Fellow, IEEE*

***Abstract*—We consider the problem of anomaly detection among a large number of processes, where the probabilistic models of anomalies are unknown. At each time, aggregated noisy observations can be taken from a chosen subset of processes, where the chosen subset conforms to a tree structure. The observation distribution depends on the chosen subset and the absence/presence of anomalies. We develop a sequential search strategy using a hierarchical Kolmogorov-Smirnov (KS) statistics. Referred to as Tree-based Anomaly Search using KS statistics (TASKS), the proposed strategy is order-optimal with respect to the size of the search space and the detection accuracy.**

***Index Terms*—Anomaly detection, dynamic search, sequential design of experiments.**

I. Introduction

HIERARCHICAL search algorithms provide an effective approach for a quick and reliable inference of abnormal behaviour, including applications in financial transactions [1], computer vision [2], and computer and communication networking [3], [4]. A common model is a tree structure, which allows easy aggregation of data flows (e.g., based on their IP-prefix). In this work, we develop a sequential search strategy for detecting unknown anomalies in massive data streams based on noisy hierarchical observations (see Fig. 1). Due to the unknown distributions of anomalous processes, the problem belongs to the domain of goodness-of-fit tests [5], in which fit between samples is often measured by the Kolmogorov-Smirnov (KS) statistics [6]–[8]. Departing from the classical goodness-of-fit tests is the hierarchical observation structure, which adds intriguing complexity in terms of how to zoom in and out on the observation tree to achieve the optimal sample complexity with respect to both the detection accuracy and the size of the search space. We develop a novel sequential search strategy using a hierarchical KS statistics for reliable and efficient anomaly detection. Referred to as Tree based Anomaly Search using KS statistics (TASKS), the proposed strategy is shown to offer order-optimal sample complexity with respect to the size of the search space and the detection accuracy.

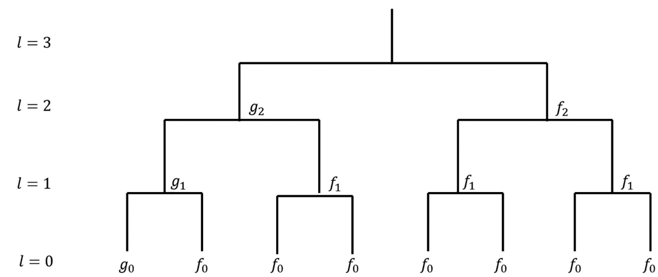

Fig. 1. A binary-tree observation model.

Hierarchical search has an intrinsic connection to the group testing problem [9]–[17]. Most existing works on group testing assume error-free test outcomes, with a few exceptions focusing on binary noisy outcomes with known noise models [18]–[20]. A recent work in [21] tackles unknown noise, but considers only discrete observations. The adaptive nature of the search strategy also shares similarities with adaptive sensing problems for sparse signal detection and support estimation [22]–[24], and to the pure-exploration bandit problems [25], [26]. Using KS statistics, the algorithm proposed in this work is fundamentally different from these existing results.

The active search problem is also within the umbrella of active hypothesis testing pioneered by Chernoff in [27] (with extensive follow-up works, e.g., [28]–[34]). Recently, Chernoff's framework was extended to handle the anomaly detection framework [35]–[42]. However, these studies either assumed known/parametric models or adopted a linear (i.e., non-hierarchical) search structure. The problem of detecting anomalies or outlying sequences has also been studied under different formulations, assumptions, and objectives in [43]–[50]. The recent works in [51], [52] considered hierarchical search under unknown observation models. The key difference is that the search strategies in [51], [52] are based on sample mean statistics, in contrast to the KS test statistics used in this work. We show in Sec. V the superior performance of TASKS over these existing strategies.

II. System Model and Problem Formulation

We consider the problem of detecting an anomalous process (i.e., a target) among a large number M of processes (i.e.,

Manuscript received June 20, 2021; revised August 15, 2021; accepted August 16, 2021. Date of publication August 20, 2021; date of current version September 14, 2021. The work of Tomer Gafni and Kobi Cohen was supported in part by the Cyber Security Research Center at Ben-Gurion University of the Negev, and in part by U.S.-Israel Binational Science Foundation (BSF) under Grant 2017723. The work of Qing Zhao was supported in part by the National Science Foundation under Grant CCF1815559. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. John Ball. *(Corresponding author: Tomer Gafni.)*

Tomer Gafni and Kobi Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: gafnito@post.bgu.ac.il; kobi.cohen10@gmail.com).

Qing Zhao is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: qz16@cornell.edu).

Digital Object Identifier 10.1109/LSP.2021.3106587

cells). If the target is in cell m, we say that hypothesis H_m is true. We assume a tree-structured hierarchy among the process distributions, denoted by $\mathcal{T}$, as illustrated in Fig. 1. Let (l, k) $(l = 0, \ldots, \log_2 M, k = 1, \ldots, 2^{\log_2 M - l})$ denote node k at level l of the tree, and let g_0 and f_0 denote, respectively, the distribution of the anomalous process and the normal processes. Let $g_l \quad (l = 1, \ldots, \log_2 M)$ denote the distribution of the measurement that aggregates the anomalous process and $2^l - 1$ normal processes, and $f_l \quad (l = 1, \ldots, \log_2 M)$ the distribution of the measurement that aggregates 2^l normal processes.

For all l we assume that f_l is known, but g_l is unknown. Also, the tree structure is known. Let F_l, G_l be the cumulative distribution function (CDF) of f_l, g_l, respectively. We assume that g_l satisfies:

$$\{g_l \ : \ \sup_x |F_l(x) - G_l(x)| \geq \Delta\} \quad \forall l, \tag{1}$$

meaning that the distributions are distinguishable by $\Delta > 0$ for all l, and Δ is independent of M (note that Δ is a known lower bound on the KS distances in all levels l). Let $\mathbb{P}_m$ be the probability measure under hypothesis H_m and $\mathbb{E}_m$ the operator of expectation with respect to the measure $\mathbb{P}_m$. We aim to develop an active search strategy $\Gamma = (\{a_n\}_{n=1}^{\tau-1}, \tau, \delta)$ that determines whether to terminate the search (at stopping time τ), and if not (for time $1 \leq n < \tau$), which node on the tree to probe next, defined by action $a_n \in \{l, k\}$, $l = 0, \ldots, \log_2 M, k = 1, \ldots, 2^{\log_2 M - l}$. A terminal decision rule $\delta \in \{1, 2, \ldots, M\}$ denotes the detected location of the target declared at time τ. We define the Bayes risk as follows: A sampling cost of $c \in (0, 1)$ is incurred for each observation, the loss for wrong declaration is 1, and π_m is the a-priori probability that process m is anomalous. Then, the error probability is given by:

$$P_e(\Gamma) \triangleq \sum_{m=1}^{M} \pi_m \mathbb{P}_m(\delta \neq m | \Gamma), \tag{2}$$

the sample complexity $\mathbb{E}[\tau|\Gamma]$ is given by:

$$\mathbb{E}[\tau|\Gamma] \triangleq \sum_{m=1}^{M} \pi_m \mathbb{E}_m[\tau|\Gamma], \tag{3}$$

and the Bayes risk is defined as:

$$R(\Gamma) \triangleq P_e(\Gamma) + c \cdot \mathbb{E}[\tau|\Gamma]. \tag{4}$$

III. The Tree-Based Anomaly Search Using KS Statistics (TASKS) Algorithm

The TASKS algorithm is executed in two interleaving phases as described next.

The inference phase: This phase is carried out on a specific node (random process) $\{X(n)\}_{n=1}^{\infty}$ on level l. The goal is to quantify the degree of normality of the node using the KS statistics. We take a fixed number of samples N_l from the node, and generate the empirical CDF:

$$\hat{F}_l^{N_l}(x) = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathbb{1}_{[-\infty, x]}(X_n), \tag{5}$$

where $\mathbb{1}_{[-\infty, x]}(X_n)$ is the indicator function, equals to 1 if $X_n \leq x$ and 0 otherwise. Next, we derive the KS statistics which quantifies the distance between the empirical CDF from the reference distribution:

$$D_{N_l} = \sup_x |\hat{F}_l^{N_l}(x) - F_l(x)|. \tag{6}$$

A high value of D_{N_l} indicates a higher probability that the node is anomalous, and vice versa. The choice of N_l should ensure that the search phase is more likely to move towards the target with a desired probability, as we discuss next.

The search phase: Based on the output of the inference phase, we specify the searching strategy of the TASKS algorithm. We start from the root, and for each stage of the walk at level $l \geq 1$ we formulate a ternary hypothesis testing problem– $\widetilde{H}_0$ refers to that the currently sampled node does not contain the target; $\widetilde{H}_1$ refers to that the left child of this node contains the target, and $\widetilde{H}_2$ refers to that the right child of the node contains the target.

Suppose that we test node i at level $l \geq 1$. We compute $D_{N_{l-1}}$ for the left and right child of the node, as described in the inference phase. If $D_{N_{l-1}}$ of both children is smaller then a pre-determined threshold γ_{l-1}, then hypothesis $\widetilde{H}_0$ is chosen, and we go back to the parent of node i. Otherwise, we zoom into the child node that has the larger KS statistics. On level $l > 1$, the probability to zoom correctly into a child node should be larger then 0.5, to ensure that the search phase is biased towards the leaf target. On level $l = 1$, if the KS statistics for the tested leaf is larger than a threshold γ_0, we declare that the leaf contains the target and terminate the search. Here the number of samples taken for generating the KS statistics should ensure the desired accuracy.

IV. Performance Analysis

Theorem 1: Let $N_l > \max\{\frac{1.38}{\gamma_{l-1}^2}, \frac{1.38}{(\Delta - \gamma_{l-1})^2}\}$ for $2 < l < \log_2 M$, and $N_1 > \frac{1}{2\gamma_0^2} \cdot \ln(\frac{\log_2 M}{c})$, where $0 < \gamma_l < \Delta, \forall l$, and Δ is defined in (1). Then, the Bayes risk under the TASKS algorithm is bounded by:

$$R(\Gamma_{\text{TASKS}}) \leq A \cdot c \cdot \log_2 M + B \cdot c \cdot \ln\left(\frac{\log_2 M}{c}\right) + O(c), \tag{7}$$

where A and B (given in (23)) are constants independent of M and c.

The optimality of the Bayes risk of TASKS in both c and M directly carries through to the sample complexity of TASKS. Specifically, from (7), we have the following upper bound on the sample complexity:

$$\mathbb{E}[\Gamma_{\text{TASKS}}] \leq A \cdot \log_2 M + B \cdot \ln\left(\frac{\log_2 M}{c}\right) + O(1). \tag{8}$$

Using the lower bound on the sample complexity which was developed in Theorem 2 in [33], for any policy Γ, we have:

$$\mathbb{E}[\Gamma_{\text{TASKS}}] \geq \frac{\log_2 M}{I_{\max}} + \frac{\log((1-c)/c)}{D(g_0||f_0)} + O(1), \tag{9}$$

where $I_{\max}$ denotes the maximum mutual information between the true hypothesis and the observation under an optimal action, and $D(g_0||f_0)$ is the KL divergence between g_0 and f_0. As a result, we get that TASKS is order optimal in c and M.

We note that the distributions of the aggregated observations is a function of the distributions of the children. However, this does not violate our assumptions in evaluating the performance, since at each inference phase, the TASKS algorithm collects new samples from the tested nodes. We next prove Theorem 1.

Proof: First, we analyze the sample complexity of the inference phase, and then analyze the search phase to establish the number of times that the inference phase is carried out. This yields the sample complexity of TASKS. Finally, we bound the probability of error, and get the desired bound for the Bayes risk.

Step 1: The sample complexity of the inference phase: We start by presenting two lemmas, which bound the type 1 error (rejection of a true null hypothesis) and type 2 error (non-rejection of a false null hypothesis).

Lemma 1: Given a natural number N, let $X_1, X_2, \ldots, X_N$ be real valued i.i.d r.v with CDF $F(\cdot)$. Let $\hat{F}^N(x)$ denote the associated empirical CDF as defined in (5). Then, for every $\gamma > 0$:

$$\mathbb{P}(\sup_x |\hat{F}^N(x) - F(x)| > \gamma) \leq 2e^{-2N\gamma^2} \quad (10)$$

Proof: The bound can be derived by applying the Dvoretzky-Kiefer-Wolfowitz inequality [53].

We now bound the type 2 error under the distinguishable assumption between the distributions (1).

Lemma 2: Given a natural number N, let $X_1, X_2, \ldots, X_N$ be real valued i.i.d r.v. with CDF $G(\cdot)$, and let $\hat{G}^N(x)$ denote the associated empirical CDF. Assume also that there exists a constant $\Delta > 0$ such that

$$\sup_x |F(x) - G(x)| \geq \Delta. \quad (11)$$

Then, for every $0 < \gamma < \Delta$ we have:

$$\mathbb{P}(\sup_x |\hat{G}^N(x) - F(x)| < \gamma) \leq 2e^{-2N(\Delta-\gamma)^2}. \quad (12)$$

Proof: Note that given (11), $\sup_x |\hat{G}^N(x) - F(x)| < \gamma$ implies:

$\sup_x |\hat{G}^N(x) - G(x)| = \sup_x |\hat{G}^N(x) - F(x) + F(x) - G(x)| \geq \sup_x |F(x) - G(x)| - \sup_x |\hat{G}^N(x) - F(x)| > \Delta - \gamma$,

and hence

$\{\sup_x |\hat{G}^N(x) - F(x)| < \gamma\} \subseteq \{\sup_x |\hat{G}^N(x) - G(x)| > \Delta - \gamma\}$.

Combining with (10) we have:

$\mathbb{P}(\sup_x |\hat{G}^N(x) - F(x)| < \gamma) \leq \mathbb{P}(\sup_x |\hat{G}^N(x) - G(x)| > \Delta - \gamma) \leq 2e^{-2N(\Delta-\gamma)^2}$. ∎

Based on Lemmas 1 and 2 we determine the number of samples we need to take in the inference phase in order to ensure a biased random walk towards the target. We define $p_l^{(g)}$ as the probability that we zoom in correctly to the anomalous child of a node at level l. Thus,

$$p_l^{(g)} \triangleq \mathbb{P}(\sup_x |\hat{G}_{l-1}^{N_l}(x) - F_{l-1}(x)| > \max\{\sup_x |\hat{F}_{l-1}^{N_l}(x) - F_{l-1}(x)|, \gamma_{l-1}\}), \quad (13)$$

where $\hat{G}_{l-1}^{N_l}(x)$ and $\hat{F}_{l-1}^{N_l}(x)$ are the empirical CDF for the abnormal process g_{l-1} and the normal process f_{l-1}, respectively, and $0 < \gamma_{l-1} < \Delta$ is a fixed tuning parameter. $p_l^{(f)}$ is defined as the probability that we return to the parent of the node when the node is normal, i.e. identifying correctly that both children are normal:

$$p_l^{(f)} \triangleq [\mathbb{P}(\sup_x |\hat{F}_{l-1}^{N_l}(x) - F_{l-1}(x)| < \gamma_{l-1})]^2. \quad (14)$$

At level $l > 1$, we can choose $p_l^{(g)} = p_l^{(f)} > 0.5$, so the random walk drifts towards the target.

We now determine the number of samples we need to take in each test. From (14) we get:

$$\sqrt{p_l^{(f)}} = \mathbb{P}(\sup_x |\hat{F}_{l-1}^{N_l}(x) - F_{l-1}(x)| < \gamma_{l-1}) = 1 - \mathbb{P}(\sup_x |\hat{F}_{l-1}^{N_l}(x) - F_{l-1}(x)| \geq \gamma_{l-1}) \geq 1 - 2e^{-2N_l\gamma_{l-1}^2},$$

where the last inequality is due to (10). To ensure $p_l^{(f)} > \frac{1}{2}$ we have $1 - 2e^{-2N_l\gamma_{l-1}^2} > \frac{1}{\sqrt{2}}$, which implies

$$N_l > \frac{0.96}{\gamma_{l-1}^2}. \quad (15)$$

Similarly, by applying Lemma 2 and some algebraic manipulations, in order to have $p_l^{(g)} > \frac{1}{2}$, we need:

$$N_l > \max\left\{\frac{1.38}{\gamma_{l-1}^2}, \frac{1.38}{(\Delta - \gamma_{l-1})^2}\right\}, \quad (16)$$

and for both (15) and (16) to hold we take:

$$N_l > \max\left\{\frac{1.38}{\gamma_{l-1}^2}, \frac{1.38}{(\Delta - \gamma_{l-1})^2}\right\}. \quad (17)$$

Finally, we define $N_{\max} = \max_{l>1}\{N_l\}$.

For the leaf nodes,[1] we should bound the probability of detection error. We design the type 1 error to be smaller than $\frac{2c}{\log_2 M}$ (as will be explained later):

$$\mathbb{P}(\sup_x |\hat{F}_0^{N_1}(x) - F_0(x)| > \gamma_0) \leq 2e^{-2N_1\gamma_0^2} \leq \frac{2c}{\log_2 M},$$

and therefore:

$$N_1 > \frac{1}{2\gamma_0^2} \cdot \ln\left(\frac{\log_2 M}{c}\right). \quad (18)$$

Step 2: Upper bound on the number of times the inference phase is called:

First, we consider a sequence of sub-trees $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{\log_2 M}$ of the tree $\mathcal{T}$. Sub-tree $\mathcal{T}_{\log_2 M}$ is obtained by removing the biggest half-tree containing the target from $\mathcal{T}$. Sub-tree $\mathcal{T}_l$ is iteratively obtained by removing the biggest half-tree containing the target from the half-tree containing the target in the previous step. For example, in Fig. 1, $\mathcal{T}_{\log_2 M} = \mathcal{T}_3$ is the sub-tree containing the four most right leaves, $\mathcal{T}_2$ is the sub-tree containing the third and fourth leaves (counted from the left), and $\mathcal{T}_1$ is the second leaf (counted from the left). Next, we consider the last passage time τ_l of the search phase from each sub-tree $\mathcal{T}_l$. We prove an upper bound on each $\mathbb{E}[\tau_l]$. For this, we define the distance of the search phase to the anomalous process as the sum of the discrete distance on the tree. The search initially starts at distance $\log_2 M$ from the target. The parameter W_n is a r.v. defined as the step of the search phase at time n. Depending on the current level l, W_n is distributed as:

$$\mathbb{P}(W_n = -1) = p_l^{(f)} \; ; \; \mathbb{P}(W_n = 1) = 1 - p_l^{(f)} \quad (19)$$

[1]The analysis of the leaf node detection can be used to analyze related linear search settings in future studies, as linear search can be viewed as a constrained model, in which only the leaf nodes can be probed.

if the node is located at a sub-tree that does not contain the target, or:

$$\mathbb{P}(W_n = -1) = p_l^{(g)} \; ; \; \mathbb{P}(W_n = 1) = 1 - p_l^{(g)} \tag{20}$$

if the node is located at a sub-tree that contains the target. Since $p_l^{(g)}, p_l^{(f)} > \frac{1}{2}$ for all $l > 1$ we have: $\mathbb{E}[W_n] = 1 - 2p_l^{(g)}$ or $1 - 2p_l^{(f)}$ which are both less then 0. In order to bound $\mathbb{E}[\tau_l]$, we first bound $\mathbb{P}(\tau_l > t)$. We first prove this for $\tau_{\log_2 M}$. Note that if the search phase is within the sub-tree $\mathcal{T}_{\log_2 M}$ at step n, we have $\sum_{s=1}^{n} W_s \geq 0$. Using Hoeffding inequality for Bernoulli distributions, we have:

$$\mathbb{P}(\tau_{\log_2 M} > t) \leq \mathbb{P}(\sup\{n \geq 1 : \textstyle\sum_{s=1}^{n} W_s \geq 0\} > t)$$
$$\leq \textstyle\sum_{n=t}^{\infty} \mathbb{P}(\sum_{s=1}^{n} W_s \geq 0) \leq \sum_{n=t}^{\infty} e^{-2n(1-2p_{\min})^2}$$
$$= \frac{e^{-2t(1-2p_{\min})^2}}{1 - e^{-2(1-2p_{\min})^2}},$$

where $p_{\min} \triangleq \min_{1<l<\log_2 M}\{p_l^{(g)}, p_l^{(f)}\}$. Based on the sum of tail probabilities we get:

$$\mathbb{E}[\tau_{\log_2 M}] = \textstyle\sum_{t=0}^{\infty} \mathbb{P}\{\tau_{\log_2 M} > t\}$$
$$\leq \sum_{t=0}^{\infty} \frac{e^{-2t(1-2p_{\min})^2}}{1 - e^{-2(1-2p_{\min})^2}} = \frac{1}{(1 - e^{-2(1-2p_{\min})^2})^2} \triangleq D.$$

From the symmetry of binary tree, it can be seen that $\mathbb{E}[\tau_l] < D$ for all $l < \log_2 M$ (since $\mathbb{E}[\tau_l]$ depends on $\{p_l^{(g)}, p_l^{(f)}\}_{l=1}^{\log_2 M - 1}$ which are bounded above by $p_{\min}$). The number of times that the inference phase is called (and applied to both children) is no bigger than $2\sum_{l=1}^{\log_2 M} \mathbb{E}[\tau_l]$, hence, for $l \geq 1$ the expected number of points visited is upper bounded by $2D \log_2 M$.

Step 3: The sample complexity of TASKS:

Finally, by summing the sample complexity over the sub-trees and the leaf node, we get:

$$\mathbb{E}[\tau] \leq 2N_{\max} D \left(\log_2 M - 1\right) + N_1. \tag{21}$$

It remains to bound the probability of detection error. The number of times of visiting non-target leaf nodes is upper bounded by $2 \cdot D \cdot \log_2 M$. As discussed above, (18) ensures that the type 1 error in the leaf level is upper bounded by $\frac{2c}{\log_2 M}$. Thus,

$$\mathbb{P}_m(\delta \neq m|\Gamma) \leq 2 \cdot D \cdot \log_2 M \cdot \frac{2c}{\log_2 M} = \frac{4}{D} \cdot c = O(c). \tag{22}$$

Finally, we choose:

$$A = 2N_{\max} \cdot D, \quad B = \frac{1}{2\gamma_0^2} \tag{23}$$

and (7) holds. □

Finally, we discuss the considerations for choosing $\{N_l\}$ and $\{\gamma_l\}$. $\{N_l\}$ captures the trade-off between the sample complexity of the inference phase (increases when $\{N_l\}$ increases), and the number of times that the inference phase is called (decreases when $\{N_l\}$ increases). The thresholds $\{\gamma_l\}$ capture the trade-off between $p_l^{(f)}$ (increases when $\{\gamma_l\}$ decreases) and $p_l^{(g)}$ (decreases when $\{\gamma_l\}$ decreases). In the next section we provide the specific values that we chose for these parameters in the numerical experiments.

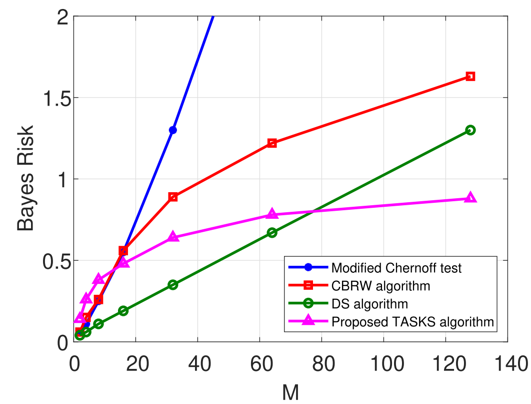

Fig. 2. The Bayes risk as a function of M ($c = 10^{-2}$).

V. Simulation Results

We simulated the case where the aggregated flows follow exponential distributions with the parameters equal to the sum of the parameters of their children at the leaf level ($\lambda = 0.1$), and a Bernoulli random interference $Z \in \{-6, 10\}$ with equal probabilities is present in the measurements that aggregate the anomalous leaf. In Fig. 2, we compared the following algorithms: (i) The Chernoff test [27] (with $\Theta_0 = \{0\}, \Theta_1 = \{10, 5, 1\}$), (ii) the DS algorithm [38] (with the same parameters as in (i)), (iii) the CBRW algorithm [51] (with $\alpha = \beta = 0.2, \xi = 0.05, \eta = 1$), and (iv) the proposed TASKS algorithm (with $N_l = 5, l = 2, 3, \ldots, \log_2 M$, $N_1 = 10$, $\gamma_l = \sqrt{1.38/N_l}$.). For all the simulations we used 1000 Monte-Carlo rounds.

We point out that even with the hierarchical observations available to the Chernoff test, it reduces to probing the leaf nodes only. The DS algorithm improves the Chernoff test by judiciously allocating exploration and exploitation phases when searching over the leaf nodes. The CBRW algorithm exploits the hierarchical structure of the flow aggregation to obtain a logarithmic search order, similarly to TASKS. It can be seen that when the number of processes is small ($M < 80$) the DS algorithm performs the best. However, as M increases the optimal logarithmic order of TASKS dominates the search, and TASKS significantly outperforms all other algorithm, including the logarithmic rate of CBRW. While CBRW was shown to be order optimal in [51], the above simulation results show that when anomaly is not prominently reflected in a mean deviation, the finite-time performance of CBRW is inferior to TASKS.

VI. Conclusion

We developed a novel sequential search strategy for the hierarchical non-parametric anomaly detection problem, dubbed TASKS. It uses the Kolmogorov-Smirnov statistics to design a biased random walk for a quick detection of the anomaly process. TASKS is shown to be order-optimal with respect to the size of the search space and the detection accuracy.

References

[1] D.-P. Li, S.-J. Cheng, P.-F. Cheng, J.-Q. Wang, and H.-Y. Zhang, "A novel financial risk assessment model for companies based on heterogeneous information and aggregated historical data," *PLoS One*, vol. 13, no. 12, 2018, Art. no. e0208166.

[2] U. Qidwai, M. Akbar, M. Maqbool, and M. Jahanshahi, "Hierarchical inspection system using visual and MFL probe robots," *Int. J. Robot. Automat.*, vol. 7, no. 4, pp. 283–296, 2018.

[3] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava, "Finding hierarchical heavy hitters in data streams," in *Proc. VLDB Conf.*, 2003, pp. 464–475.

[4] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," *IEEE Netw.*, vol. 11, no. 6, pp. 10–23, Nov./Dec. 1997.

[5] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques*, Boca Raton, FL, USA: CRC Press, 1986.

[6] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[7] F. J. Massey Jr., "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, pp. 68–78, 1951.

[8] D. Quade, "On the asymptotic power of the one-sample Kolmogorov-Smirnov tests," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1000–1018, 1965.

[9] D. Sejdinovic and O. Johnson, "Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 998–1003.

[10] J. Scarlett and V. Cevher, "Converse bounds for noisy group testing with arbitrary measurement matrices," in *Proc. IEEE Int. Symp. Inf. Theory*, 2016, pp. 2868–2872.

[11] J. Scarlett and V. Cevher, "Near-optimal noisy group testing via separate decoding of items," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 902–915, Oct. 2018.

[12] H. A. Inan, P. Kairouz, and A. Ozgur, "Energy-limited massive random access via noisy group testing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 1101–1105.

[13] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, Mar. 2012.

[14] V. Y. Tan and G. K. Atia, "Strong impossibility results for noisy group testing," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 8257–8261.

[15] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Grotesque: Noisy group testing (quick and efficient)," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, 2013, pp. 1234–1241.

[16] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 3019–3035, May 2014.

[17] Y. Kaspi, O. Shayevitz, and T. Javidi, "Searching with measurement dependent noise," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2690–2705, Apr. 2018.

[18] D. Malioutov and M. Malyutov, "Boolean compressed sensing: LP relaxation for group testing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 3305–3308.

[19] K. Lee, K. Chandrasekher, R. Pedarsani, and K. Ramchandran, "Saffron: A fast, efficient, and robust framework for group testing based on sparse-graph codes," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4649–4664, Sep. 2019.

[20] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3646–3661, Jun. 2018.

[21] S. Salgia and Q. Zhao, "An order-optimal adaptive test plan for noisy group testing under unknown noise models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 4035–4039.

[22] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20, no. 4, pp. 2217–2246, 2014.

[23] J. Haupt, R. M. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6222–6235, Sep. 2011.

[24] M. A. Davenport and E. Arias-Castro, "Compressive binary search," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 1827–1831.

[25] S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen, "Combinatorial pure exploration of multi-armed bandits," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 379–387, 2014.

[26] A. Locatelli, M. Gutzeit, and A. Carpentier, "An optimal algorithm for the thresholding bandit problem," in *Proc. Int. Conf. Mach. Learn., PMLR*, 2016, pp. 1690–1698.

[27] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, vol. 30, no. 3, pp. 755–770, 1959.

[28] S. A. Bessler, "Theory and applications of the sequential design of experiments, K-actions and infinitely many experiments. Part I. Theory," App. Math. Statistics Labs, Stanford Univ., Tech. Rep., 1960.

[29] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *IEEE Trans. Autom. Control*, vol. 58, no. 10, pp. 2451–2464, Oct. 2013.

[30] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for hypothesis testing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 5277–5280.

[31] M. Naghshvar and T. Javidi, "Sequentiality and adaptivity gains in active hypothesis testing," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 768–782, Oct. 2013.

[32] S. Nitinawarat and V. V. Veeravalli, "Controlled sensing for sequential multihypothesis testing with controlled Markovian observations and non-uniform control cost," *Sequential Anal.*, vol. 34, pp. 1–24, 2015.

[33] M. Naghshvar *et al.*, "Active sequential hypothesis testing," *Ann. Statist.*, vol. 41, no. 6, pp. 2703–2738, 2013.

[34] D. Kartik, E. Sabir, U. Mitra, and P. Natarajan, "Policy design for active sequential hypothesis testing using deep learning," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2018, pp. 741–748.

[35] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1432–1450, Mar. 2015.

[36] K. Cohen and Q. Zhao, "Asymptotically optimal anomaly detection via sequential testing," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2929–2941, Jun. 2015.

[37] B. Huang, K. Cohen, and Q. Zhao, "Active anomaly detection in heterogeneous processes," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2284–2301, Apr. 2018.

[38] B. Hemo, T. Gafni, K. Cohen, and Q. Zhao, "Searching for anomalies over composite hypotheses," *IEEE Trans. Signal Process.*, vol. 68, pp. 1181–1196, 2020, doi: 10.1109/TSP.2020.2971438.

[39] A. Tsopelakos, G. Fellouris, and V. V. Veeravalli, "Sequential anomaly detection with observation control," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 2389–2393.

[40] C. Wang, K. Cohen, and Q. Zhao, "Information-directed random walk for rare event detection in hierarchical processes," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 1099–1116, Feb. 2021.

[41] A. Tsopelakos and G. Fellouris, "Sequential anomaly detection with observation control under a generalized error metric," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 1165–1170.

[42] C. Zhong, M. C. Gursoy, and S. Velipasalar, "Deep actor-critic reinforcement learning for anomaly detection," in *Proc. IEEE Glob. Commun. Conf.*, 2019, pp. 1–6.

[43] J. Heydari, A. Tajer, and H. V. Poor, "Quickest linear search over correlated sequences," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5786–5808, Oct. 2016.

[44] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4462–4481, Jul. 2013.

[45] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 44–56, Sep. 2014.

[46] J. Geng, W. Xu, and L. Lai, "Quickest search over multiple sequences with mixed observations," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 2582–2586.

[47] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a Markov process across a sensor array," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1961–1981, Apr. 2010.

[48] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 512–525, Apr. 2011.

[49] Y. Song and G. Fellouris, "Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals," *Electron. J. Statist.*, vol. 11, no. 1, pp. 338–363, 2017.

[50] Y. Song and G. Fellouris, "Sequential multiple testing with generalized error control: An asymptotic optimality theory," *Ann. Statist.*, vol. 47, no. 3, pp. 1776–1803, 2019.

[51] S. Vakili, Q. Zhao, C. Liu, and C.-N. Chuah, "Hierarchical heavy hitter detection under unknown models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6917–6921.

[52] S. Vakili and Q. Zhao, "A random walk approach to first-order stochastic convex optimization," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 395–399.

[53] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statist.*, 1956, pp. 642–669, 1956.