Optimal Index Policies for Anomaly Localization in Resource-Constrained Cyber Systems

Kobi Cohen[1], Qing Zhao[1], Ananthram Swami[2]

***Abstract*— The problem of anomaly localization in a resource-constrained cyber system is considered. Each anomalous component of the system incurs a cost per unit time until its anomaly is identified and fixed. Different anomalous components may incur different costs depending on their criticality to the system. Due to resource constraints, only one component can be probed at each given time. The observations from a probed component are realizations drawn from two different distributions depending on whether the component is normal or anomalous. The objective is a probing strategy that minimizes the total expected cost, incurred by all the components during the detection process, under reliability constraints. We consider both independent and exclusive models. In the former, each component can be abnormal with a certain probability independent of other components. In the latter, one and only one component is abnormal. We develop optimal simple index policies under both models. The proposed index policies apply to a more general case where a subset (more than one) of the components can be probed simultaneously and have strong performance as demonstrated by simulation examples.**

***Index Terms*—** Anomaly localization, Sequential Probability Ratio Test (SPRT), sequential hypothesis testing, detection under uncertainty.

I. Introduction

We consider anomaly localization where the objective is to identify anomalous components in a system quickly and reliably. Consider a cyber system with K components. Each component may be in a normal or an abnormal state. If abnormal, component k incurs a cost c_k per unit time until its anomaly is identified and fixed. Due to resource constraints, only one component can be probed at a time, and switching to a different component is allowed only when the state of the current component is declared. The observations from a probed component (say k) follow distributions $f_k^{(0)}$ or $f_k^{(1)}$ depending on whether the component is normal or anomalous, respectively. The objective is a probing strategy that dynamically determines the order of the sequential tests performed on all the components so that the total cost incurred to the system during the entire detection process is minimized under reliability constraints.

[1] Department of Electrical and Computer Engineering, University of California, Davis. Email: {yscohen, qzhao}@ucdavis.edu.

[2] Army Research Laboratory, Adelphi, MD 20783. Email: a.swami@ieee.org.

This work was supported by Army Research Lab under Grant W911NF1120086.

Part of this work will be presented at the IEEE Global Conference on Signal and Information Processing (GlobalSIP), Austin, Texas, USA, Dec. 2013.

A. Main Results

The above problem presents an interesting twist to the classic sequential hypothesis testing problem. In the case when there is only one component, minimizing the cost is equivalent to minimize the detection delay, and the problem is reduced to a classic sequential test where both the simple and the composite hypothesis cases have been well studied. With multiple components, however, minimizing the detection delay of each component is no longer sufficient. The key to minimize the total cost is the order at which the components are being tested. It is intuitive that we should prioritize components with higher costs when abnormal and components with higher prior probabilities for being abnormal. Another parameter that plays a role in the total system cost is the expected time in detecting the state of a component which depends on the observation distributions $\{f_k^{(0)}, f_k^{(1)}\}$: it is desirable to place components that require longer testing time toward the end of the testing process. The challenge here is how to balance these key parameters in the dynamic probing strategy.

We show in this paper that the optimal probing strategy is an open-loop policy where the testing order can be predetermined, independent of the realizations of each individual test in terms of both the test outcome and the detection time. Furthermore, the probing order is given by a simple index. Specifically, under the independent model where each component is abnormal with probability π_k independent of other components, the index is in the form of $\pi_k c_k/\mathbf{E}(N_k)$, where $\mathbf{E}(N_k)$ is the expected detection time for component k. Under the exclusive model where one and only one component is abnormal, the index is in the form of $\pi_k c_k/\mathbf{E}(N_k|H_0)$ where $\mathbf{E}(N_k|H_0)$ is the expected detection time for component k under the hypothesis of it being normal. It is interesting to notice the difference in the indexes for these two models. Intuitively speaking, under the exclusive model, the detection times of the *normal* components tested before the single abnormal one add to the cost incurred by the abnormal component, while under the independent model, the detection time of any component, normal or abnormal, adds to the delay in catching the next abnormal component.

The above simple index forms of the probing order are optimal for both the simple hypothesis ($\{f_k^{(0)}, f_k^{(1)}\}_{k=1}^K$ are known) and the composite hypothesis ($\{f_k^{(0)}, f_k^{(1)}\}_{k=1}^K$ have unknown parameters) cases. These index policies also apply to the case where more than one component can be probed simultaneously and offer strong performance as demonstrated by simulation examples. Their optimality in this case, however, remains open.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Optimal Index Policies for Anomaly Localization in Resource-Constrained Cyber Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Davis,Department of Electrical and Computer Engineering,Davis,CA,95616**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
submitted to IEEE Transactions on Signal Processing, October, 2013.

14. ABSTRACT
The problem of anomaly localization in a resourceconstrained cyber system is considered. Each anomalous component of the system incurs a cost per unit time until its anomaly is identified and fixed. Different anomalous components may incur different costs depending on their criticality to the system. Due to resource constraints, only one component can be probed at each given time. The observations from a probed component are realizations drawn from two different distributions depending on whether the component is normal or anomalous. The objective is a probing strategy that minimizes the total expected cost incurred by all the components during the detection process under reliability constraints. We consider both independent and exclusive models. In the former, each component can be abnormal with a certain probability independent of other components. In the latter, one and only one component is abnormal. We develop optimal simple index policies under both models. The proposed index policies apply to a more general case where a subset (more than one) of the components can be probed simultaneously and have strong performance as demonstrated by simulation examples.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**13**	

B. Applications

In addition to anomaly detection in cyber systems, the above problem also finds applications in spectrum scanning in cognitive radio systems and event detection in sensor networks. In the following we give two specific examples.

Consider a cyber network consisting of K components (which can be routers, paths, etc.). Due to resource constraints, only a subset of the components can be probed at a time. An Intrusion Detection System (IDS) analyzes the traffic over the components to detect Denial of Service (DoS) attacks (such attacks rely on overwhelming the component with useless traffic that exceeds its capacity so as to make it unavailable for its intended use). Let the cost c_k be the normal expected traffic (packets per unit time) over component k. Thus, in this example minimizing the total expected cost minimizes the total expected number of failed packets in the network during DoS attacks. The exclusive model applies to cases where an intrusion to a subnet, consisting of K components, has been detected and the probability of each component being compromised is small (thus with high probability, there is only one abnormal component).

Another example is spectrum sensing in cognitive radio systems. Consider a spectrum consisting of K orthogonal channels. Accessing an idle channel leads to a successful transmission, while accessing a busy channel results in a collision with other users. A Cognitive Radio (CR) is an intelligent device that can detect and access idle channels in the wireless spectrum. Due to resource constraints, only a subset of the channels can be sensed at a time. Once a channel is identified as idle the CR transmits over it. Let c_k be the achievable rate over channel k. Thus, in this example minimizing the total expected cost minimizes the total expected loss in data rate during the spectrum sensing process.

C. Related Work

The anomaly localization problem, studied in this paper, presents an interesting twist to the classic sequential hypothesis testing problem which considers only a single stochastic process. Sequential hypothesis testing was pioneered by Wald [1]. Wald derived the Sequential Probability Ratio Test (SPRT) for a binary hypothesis testing. Under the simple hypotheses case, the SPRT is optimal in terms of minimizing the expected sample size under given type I and type II error probability constraints. Various extensions for M-ary hypothesis testing and testing composite hypotheses were studied in [2]–[8] for a single process. In these cases, asymptotically optimal performance can be obtained as the error probability approaches zero.

Differing from this work, most of the existing studies on sequential detection over multiple processes focus on minimizing the total detection delay. Sequential detection over independent processes have been considered in [9]–[14]. In [9], [10], the problem of quickly detecting an idle period over multiple independent ON/OFF processes was considered. An optimal threshold policy was derived in [10]. The ON/OFF nature of the processes and the objective of minimizing the total detection delay make the problems considered in [9], [10] fundamentally different from the one considered in this work. In [11], the problem of quickest detection of idle channels over K independent channels with fixed idle/busy state was studied. The objective is to minimize the detection delay under error constraints. It was shown that the optimal policy is to carry out an independent SPRT over each channel, irrespective of the testing order. In contrast to [11], we show in this paper that the optimal policy in our model highly depends on the testing order even when the processes are independent. In [12], the problem of identifying the first abnormal sequence among an infinite number of i.i.d sequences was considered. An optimal cumulative sum (CUSUM) test was established under this setting. The sequential search problem under the exclusive model was investigated in [15]–[18]. Optimal policies were derived for the problem of quickest search over Weiner processes [15]–[17]. It was shown in [15], [16] that the optimal policy is to select the sequence with the highest posterior probability of being the target at each given time. In [17], an SPRT-based solution was derived, which is equivalent to the optimal policy in the case of searching over Weiner processes. However, minimizing the total expected cost in our model leads to a different problem and consequently a different index policy.

The classic target whereabouts problem is also a detection problem over multiple processes. In this problem, multiple locations are searched to locate a target. The problem is often considered under the setting of fixed sample size as in [19]–[22]. In [19], [20], [22], searching in a specific location provides a binary-valued measurement regarding the presence or absence of the target. In [21], Castanon considered the dynamic search problem under continuous observations: the observations from a location without the target and with the target have distributions f and g, respectively. The optimal policy was established under a symmetry assumption that $f(x) = g(b - x)$ for some b.

The anomaly detection problem can be considered as a special case of active hypothesis testing in which the decision maker chooses and dynamically changes its observation model among a set of observation options. Classic and more recent studies of general active hypothesis testing problems can be found in [23]–[27].

D. Organization

In Section II we describe the system model and problem formulation. In Section III we propose a two-stage optimization problem that simplifies computation while preserving optimality. In Section IV we derive optimal algorithms under the independent and exclusive models for the simple hypotheses case. In Section V we extend our results to the composite hypothesis case: we derive asymptotically optimal algorithms under the independent and exclusive models. In Section VI we provide numerical examples to illustrate the performance of the algorithms.

II. System Model and Problem Formulation

Consider a cyber system consisting of K components, where every component may be in a normal state (i.e., healthy)

or abnormal state. Define

$$\begin{aligned}\mathcal{H}_1 &\triangleq \{k : 1 \leq k \leq K \text{ , component } k \text{ is abnormal}\} \text{ ,} \\ \mathcal{H}_0 &\triangleq \{k : 1 \leq k \leq K \text{ , component } k \text{ is healthy}\} \text{ ,}\end{aligned} \tag{1}$$

as the sets of the abnormal and healthy components.
We consider two different anomaly models.

1) Exclusive model: One and only one component is abnormal with *a priori* probability π_k, where $\sum_{k=1}^{K} \pi_k = 1$.

2) Independent model: Each component k is abnormal with *a priori* probability π_k independent of other components.

Every abnormal component k incurs a cost c_k $(0 \leq c_k < \infty)$ per unit time until it is tested and identified. Components in a normal state do not incur cost. We focus on the case where only one component can be probed at a time. The resulting probing strategies apply to the case where a subset of the components can be probed simultaneously and their performance in this case are studied via simulation examples, given in Sec. VI. When component k is tested at time t, a measurement (or a vector of measurements) $y_k(t)$ is drawn independently in a one-at-a-time manner. If component k is healthy, $y_k(t)$ follows distribution $f_k^{(0)}$; if component k is abnormal, $y_k(t)$ follows distribution $f_k^{(1)}$. We focus first on the simple hypotheses case, where the distributions $f_k^{(0)}$, $f_k^{(1)}$ are completely known. In Section V we extend our results to the composite hypotheses case, where there is uncertainty in the distribution parameters.

Let $k^*(t)$ denotes the component index which is tested at time t. Let $\mathbf{y}(t) = \left\{k^*(i), y_{k^*(i)}\right\}_{i=1}^{t}$ be the set of all the available observations (and the component indices) up to time t. A selection rule is a mapping from $\mathbf{y}(t-1)$ to $\{1, 2, ..., K\}$, which indicates which component is chosen to be tested at time t. A stopping rule and a decision rule are used to decide when to terminate the test and which components are declared as abnormal, respectively.

Remark 1: Computing optimal policies for detection problems involving multiple sequences becomes impractical in general as the number of sequences or the sample size increases [21]. In [15], [19]–[21], restrictive assumptions on the distributions were used to obtain simple optimal index policies. In [12], [17], [18], restrictive assumptions on the search model were used to make the problem mathematically tractable. Here, we use similar assumptions on the search model to obtain a mathematically tractable optimization problem.

We consider the case where switching between components is allowed only when the state of the current component is declared (i.e., switching without memory). From a system perspective, the advantages of this scheme are twofold. First, switching between components typically adds a significant delay that should be avoided. Second, the decision maker stores observations of only one component at each time. Thus, this scheme is applicable to limited-memory systems. For convenience, we define t_m as the time when the decision maker starts the m^{th} test. Let $\phi(t_m) \in \{1, 2, ..., K\}$ be a selection rule that indicates which component is probed at time t_m. The vector of selection rules for the K components is denoted by $\boldsymbol{\phi} = (\phi(t_1), ..., \phi(t_K))$. Let $\mathbf{1}_k(t_m)$ be the probing indicator function, where $\mathbf{1}_k(t_m) = 1$ if component k is probed at time t_m and $\mathbf{1}_k(t_m) = 0$ otherwise.

Let τ_k be a stopping time (or a stopping rule), which is the time when the the decision maker stops taking observations from component k and declares its state. The vector of stopping times for the K components is denoted by $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)$. The random sample size required to make a decision regarding the state of component k is denoted by N_k. For example, if the decision maker tests component 1 followed by component 2, then $\tau_1 = N_1$ and $\tau_2 = N_1 + N_2$. Let $\delta_k \in \{0, 1\}$ be a decision rule, which the decision maker uses to declare the state of component k at time τ_k. $\delta_k = 0$ if the decision maker declares that component k is in a healthy state (i.e., H_0), and $\delta_k = 1$ if the decision maker declares that component k is in an abnormal state (i.e., H_1). The vector of decision rules for the K components is denoted by $\boldsymbol{\delta} = (\delta_1, ..., \delta_K)$. An admissible strategy s is a sequence of K sequential tests for the K components and denoted by the tuple $s = (\boldsymbol{\tau}, \boldsymbol{\delta}, \boldsymbol{\phi})$.

The problem is to find a strategy s that minimizes the total expected cost, incurred by all the abnormal components until declaring their states, subject to type I (false-alarm) and type II (miss-detect) error constraints for each component:

$$\begin{aligned}\inf_{s} \quad & \mathbf{E}\left\{\sum_{k\in\mathcal{H}_1} c_k \tau_k\right\} \\ s.t. \quad & P_k^{FA} \leq \alpha_k \text{ , } P_k^{MD} \leq \beta_k \quad \forall k = 1, ..., K \text{ ,}\end{aligned} \tag{2}$$

Applying type I and type II error constraints for every component was done in [11] for the problem of quickest spectrum scanning over K independent channels. In this case, the optimal solution is a sequence of SPRTs (irrespective of the testing order) for the K channels [11]. However, in this paper we show that minimizing the total expected cost leads to different solutions.

Remark 2: Note that the definition of (2) does not include the cost due to missed-detection events of abnormal components. However, the probability of missed-detection events decreases exponentially with the sample size [1], [24]. Since the error probability is typically required to be small, (2) well approximates the actual loss in practice.

Remark 3: In contrast to the case of minimizing the total delay, in our model sampling normal components after all the abnormal components have been identified does not incur cost. Therefore, applying a sequence of K sequential tests with type I and type II error constraints (2) for every component is reasonable for both independent and exclusive models (note that in our model the decision maker is allowed to declare more than one component as abnormal under the exclusive model). From a system perspective, this formulation makes the scheme robust against mistakes in the system model (for instance, if an exclusive model is assumed, but there is more than one abnormal component in the system).

We develop optimal and asymptotically optimal algorithms to solve (2) under the simple and composite hypotheses cases, respectively. We show that the optimal probing strategy follows a simple index rule and is predetermined at time t_1

(i.e., open-loop) under both the independent and exclusive models.

III. Decoupling of Ordering and Sequential Testing

In this section, we show that the probing order and the sequential testing of each component can be decoupled. As a consequence, the solution to (2) can be obtained in two stages. At the first stage, the problem is to find a stopping rule τ_k and a decision rule δ_k for every component k that minimize the expected sample size given H_i subject to error probability constraints:

$$\begin{aligned} &\inf_{\tau_k,\delta_k} \mathbf{E}(N_k|H_i)\ ,\ \ i=0,1 \\ &s.t. \quad P_k^{FA} \leq \alpha_k\ \ ,\ \ P_k^{MD} \leq \beta_k\ . \end{aligned} \tag{3}$$

For the simple hypotheses case, the solution to the first-stage optimization problem (3) is given by the SPRT [1].
Assume that component k is tested at time $t=1$. Let

$$L_k(n) = \frac{\prod_{i=1}^{n} f_k^{(1)}(y_k(i))}{\prod_{i=1}^{n} f_k^{(0)}(y_k(i))} \tag{4}$$

be the Likelihood Ratio (LR) between the two hypotheses for component k at stage n.
Let A_k, B_k ($B_k > 1/A_k$) be the boundary values used by the SPRT for component k, such that the error constraints are satisfied. In the SPRT algorithm, at each stage n, the LR is compared to the boundary values as follows:

- If $L_k(n) \in \left((A_k)^{-1}, B_k\right)$, continue to take observations from component k.
- If $L_k(n) \geq B_k$, stop taking observations from component k and declare it as abnormal (i.e., $\delta_k = 1$). Clearly, $N_k = n$.
- If $L_k(n) \leq (A_k)^{-1}$, stop taking observations from component k and declare it as normal (i.e., $\delta_k = 0$). Clearly, $N_k = n$.

Remark 4: Implementation of the SPRT requires computation of A_k and B_k ensuring the constraints on the error probability. In general, the exact determination of the boundary values is very laborious and depends on the observation distribution. Wald's approximation can be applied to simplify the computation [1]:

$$B_k \approx \frac{1-\beta_k}{\alpha_k}\ \ , A_k \approx \frac{1-\alpha_k}{\beta_k}\ . \tag{5}$$

Wald's approximation performs well for small α_k, β_k. Since type I and type II errors are typically required to be small, Wald's approximation is widely used in practice [1].

At the second stage, the problem is to find a selection rule $\boldsymbol{\phi}$ that minimizes the objective function, given the solution to the K subproblems (3):

$$\inf_{\boldsymbol{\phi}}\ \mathbf{E}\left\{ \sum_{k\in\mathcal{H}_1} c_k\tau_k \ \mid (\boldsymbol{\tau}^*, \boldsymbol{\delta}^*) \right\} \tag{6}$$

where

$$\boldsymbol{\tau}^* = (\tau_1^*, ..., \tau_K^*)\ ,\ \boldsymbol{\delta}^* = (\delta_1^*, ..., \delta_K^*) \tag{7}$$

denote the vectors of stopping times and decision rules, respectively, that solve the K subproblems (3).
The solutions to the second-stage optimization problem for the independent and exclusive models are given in Section IV.

The formulation of the two-stage optimization problem allows us to decompose the original optimization problem (2) into $K+1$ subproblems (3) and (6). In subsequent sections we show that the two-stage optimization problem preserves optimality under both the independent and exclusive models.

IV. The Simple Hypotheses Case

In this section we derive optimal solutions to both the independent and exclusive models when the observation distributions under both hypotheses are completely known. Under the independent model, the posterior probability of component k being abnormal can be updated at time t_{m+1} as follows:

$$\begin{aligned} \pi_k(t_{m+1}) &= (1-\mathbf{1}_k(t_m))\,\pi_k(t_m) \\ &+ \frac{\mathbf{1}_k(t_m)\pi_k(t_m) f_k^{(1)}(\mathbf{y}_k(N_k))}{\pi_k(t_m) f_k^{(1)}(\mathbf{y}_k(N_k)) + (1-\pi_k(t_m))\, f_k^{(0)}(\mathbf{y}_k(N_k))}\ , \end{aligned} \tag{8}$$

where $\pi_k(t_1) = \pi_k$ denotes the *a priori* probability of component k being abnormal. The term $\mathbf{y}_k(N_k) = \{y_k(i)\}_{i=t_m}^{t_m+N_k-1}$ denotes the N_k-size vector of observations, taken from component k.
Under the exclusive model, $\pi_k(t_{m+1})$ is given in (9) at the top of the next page. Note that in contrast to the independent model, under the exclusive model the beliefs of all the components are changed at each time due to the dependency across components. The posterior probabilities depend on the selection rule and the collected measurements.

A. Optimal Index Policies

Based on the solution to the two-stage optimization problem, we propose Algorithms $1, 2$, presented in Tables I, II, to solve (2). In [28], the problem of ordering operations (or components) with a given processing time was considered. It was shown that the optimal selection rule for the problem of minimizing an expected weighted sum of completion times is to select the components in decreasing order of $c_k/\mathbf{E}(N_k)$. However, the problem in (6) is different. First, the components may be normal or abnormal and the expected sample size depends on the component state. Second, the objective is to minimize an expected weighted sum of stopping times of abnormal components only. Third, under the exclusive model, the state of each component depends on other components. Furthermore, the original optimization (2) is also over the stopping rules which control the expected sample size. Here, we derive optimal selection rules that solve the second-stage optimization problem (6) for the independent and exclusive models. These selection rules are given in step 1 in Tables I, II for the independent and exclusive models, respectively. Arranging the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}(N_k)$ or $\pi_k(t_1)c_k/\mathbf{E}(N_k|H_0)$ in step 1 can be done in $O(K\log K)$ time via sorting algorithms. Next, by the optimal solution to (3), a series of SPRTs is performed according to this order until all the components are tested.

$$\pi_k(t_{m+1}) = \frac{\mathbf{1}_k(t_m)\pi_k(t_m)f_k^{(1)}(\mathbf{y}_k(N_k))}{\pi_k(t_m)f_k^{(1)}(\mathbf{y}_k(N_k)) + (1-\pi_k(t_m))\, f_k^{(0)}(\mathbf{y}_k(N_k))} + \frac{(1-\mathbf{1}_k(t_m))\,\pi_k(t_m)f_{\phi(t_m)}^{(0)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)}))}{\pi_{\phi(t_m)}(t_m)f_{\phi(t_m)}^{(1)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)})) + \left(1-\pi_{\phi(t_m)}(t_m)\right) f_{\phi(t_m)}^{(0)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)}))} \,. \tag{9}$$

TABLE I
ALGORITHM 1 FOR THE INDEPENDENT MODEL

1.	arrange the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}(N_k)$
2.	for $k = 1, ..., K$ components do:
3.	perform SPRT for component k, with $P_k^{FA} \leq \alpha_k$, $P_k^{MD} \leq \beta_k$

TABLE II
ALGORITHM 2 FOR THE EXCLUSIVE MODEL

1.	arrange the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}(N_k\|H_0)$
2.	for $k = 1, ..., K$ components do:
3.	perform SPRT for component k, with $P_k^{FA} \leq \alpha_k$, $P_k^{MD} \leq \beta_k$

The index policies, described in Algorithms $1, 2$, are intuitively satisfying. The priority of component k in terms of testing order should be higher as the cost c_k increases, or the *a priori* probability of being abnormal $\pi_k(t_1)$ increases. Under the independent model, the priority of component k in terms of testing order should be higher as the expected sample size $\mathbf{E}(N_k)$ decreases (since $\mathbf{E}(N_k)$ contributes to the cost of every component which is tested after component k). On the other hand, under the exclusive model, the priority of component k in terms of testing order should be higher as $\mathbf{E}(N_k|H_0)$ decreases. Note that under the exclusive model, we take into account the expected sample size under H_0 solely. The reason is that if component k is abnormal, there is no additional cost, incurred by other components (since only one component is abnormal). On the other hand, if component k is healthy, then $\mathbf{E}(N_k|H_0)$ contributes to the cost of the components which are tested after component k (and may be abnormal). The SPRT is used in both models to minimize the expected sample size to reduce the total cost.

The optimality of Algorithms $1, 2$ is shown in the following theorem.

Theorem 1: Under the independent and exclusive models, Algorithms $1, 2$*, respectively, solve the original optimization problem (2).*

Proof: See Appendices VIII-A and VIII-B. ∎

Note that Algorithms $1, 2$ use open-loop selection rules (as stated in step 1), where the components order is predetermined at time t_1. However, Theorem 1 is not restricted to open-loop selection rules. Theorem 1 shows that Algorithms $1, 2$ are optimal among the class of both open-loop and closed-loop selection rules.

B. Computing the Index

Arranging the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}(N_k)$ or $\pi_k(t_1)c_k/\mathbf{E}(N_k|H_0)$ requires one to compute the expected sample size $\mathbf{E}(N_k|H_i)$ for all $k = 1, 2, ..., K$. In general, it is difficult to obtain a closed-form expression for $\mathbf{E}(N_k|H_i)$. However, since the solution to (3) is given by the SPRT, Wald's approximation can be applied to simplify the computation [1]. For every $i, j = 0, 1$, let

$$D_k(i||j) = \mathbf{E}_i\left(\log\frac{f_k^{(i)}(y_k(1))}{f_k^{(j)}(y_k(1))}\right) \tag{10}$$

be the Kullback-Leibler (KL) divergence between the hypotheses H_i and H_j, where the expectation is taken with respect to $f_k^{(i)}$.
The expected sample size conditioned on each hypothesis is well approximated by [1]:

$$\begin{aligned}\mathbf{E}(N_k|H_0) &\approx \frac{(1-\alpha_k)\log\tilde{A}_k - \alpha_k\log\tilde{B}_k}{D_k(0||1)}\,,\\ \mathbf{E}(N_k|H_1) &\approx \frac{(1-\beta_k)\log\tilde{B}_k - \beta_k\log\tilde{A}_k}{D_k(1||0)}\,,\end{aligned} \tag{11}$$

where $\tilde{A}_k = (1-\alpha_k)/\beta_k, \tilde{B}_k = (1-\beta_k)/\alpha_k$ are the approximation to A_k, B_k, given in (5).
The expected sample size required to make a decision regarding the state of component k is given by:

$$\mathbf{E}(N_k) = \pi_k\mathbf{E}(N_k|H_1) + (1-\pi_k)\mathbf{E}(N_k|H_0)\,, \tag{12}$$

where the approximation approaches the exact expected sample size for small α_k, β_k.

V. THE COMPOSITE HYPOTHESES CASE

In the previous section we focused on the simple hypotheses case, where the distribution under both hypotheses are completely known. For this case, the SPRT was applied in Algorithms $1, 2$ to solve (3). However, in numerous cases there is uncertainty in the observation distributions.

For example, Consider a one-parameter distribution. Suppose that it is required to test $\theta_k < \theta_k^{(0)}$ against $\theta_k > \theta_k^{(1)} > \theta_k^{(0)}$. As discussed in [1], the SPRT can be applied to this problem by testing $\theta_k = \theta_k^{(0)}$ against $\theta_k = \theta_k^{(1)}$, where the boundary values are set such that the error constraints are

satisfied at $\theta_k^{(0)}, \theta_k^{(1)}$. For some important cases, such as an exponential family of distributions, this sequential test has the property that type I and type II errors are less than α_k, β_k for all $\theta_k < \theta_k^{(0)}$ and $\theta_k > \theta_k^{(1)}$, respectively. However, while the SPRT minimizes the expected sample size at $\theta_k = \theta_k^{(0)}, \theta_k^{(1)}$, it is highly sub-optimal for other values of θ, as demonstrated in Section VI. Therefore, other techniques should be considered under the composite hypotheses case.

Let $\boldsymbol{\theta}_k$ be a vector of unknown parameters of component k. The observations $\{y_k(i)\}_{i\geq 1}$ are drawn from a common distribution $f\left(y|\boldsymbol{\theta}_k\right)$, $\boldsymbol{\theta}_k \in \Theta_k$, where Θ_k is the parameter space of component k. If component k is healthy, then $\boldsymbol{\theta}_k \in \Theta_k^{(0)}$; if component k is abnormal, then $\boldsymbol{\theta}_k \in (\Theta \backslash \Theta_k^{(0)})$. Let $\Theta_k^{(0)}$, $\Theta_k^{(1)}$ be disjoint subsets of Θ_k, where $I_k = \Theta \backslash (\Theta_k^{(0)} \cup \Theta_k^{(1)}) \neq \emptyset$ is an indifference region[1]. When $\boldsymbol{\theta}_k \in I_k$, the detector is indifferent regarding the state of component k. Hence, there are no constraints on the error probabilities for all $\boldsymbol{\theta}_k \in I_k$. The hypothesis test regarding component k is to test

$$\boldsymbol{\theta}_k \in \Theta_k^{(0)} \quad \text{against} \quad \boldsymbol{\theta}_k \in \Theta_k^{(1)}.$$

Narrowing I_k has the price of increasing the sample size.
Let

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k(n) &= \arg \max_{\theta_k \in \Theta_k} f\left(\mathbf{y}_k(n)|\boldsymbol{\theta}_k\right), \\ \hat{\boldsymbol{\theta}}_k^{(i)}(n) &= \arg \max_{\theta_k \in \Theta_k^{(i)}} f\left(\mathbf{y}_k(n)|\boldsymbol{\theta}_k\right), \end{aligned} \tag{13}$$

be the Maximum-Likelihood Estimates (MLEs) of the parameters over the parameter spaces Θ_k, $\Theta_k^{(i)}$ at stage n, respectively.

In contrast to the SPRT (for the simple hypotheses case), the theory of sequential tests of composite hypotheses does not provide optimal performance in terms of minimizing the expected sample size under given error constraints. Nevertheless, asymptotically optimal performance can be obtained as the error probability approaches zero.

First, we provide an overview of existing sequential tests for composite hypotheses which are relevant to our problem. Next, we apply these techniques to solve (2).

A. Existing Sequential Tests for Composite Hypothesis Testing

The key idea is to use the estimated parameters to perform a one-sided sequential test to reject H_0 and a one-sided sequential test to reject H_1. Note that these techniques were introduced for a single process. However, in this paper we apply sequential tests for K components. Thus, we use the subscript k to denote the component index.

1) Sequential Generalized Likelihood Ratio Test (SGLRT): We refer to sequential tests that use the Generalized Likelihood Ratio (GLR) statistics as the SGLRT.
For $i = 0, 1$, let

$$L_k^{(i),GLR}(n) = \log \frac{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k(n))}{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k^{(i)}(n))} \tag{14}$$

[1]The assumption of an indifference region is widely used in the theory of sequential testing of composite hypotheses to derive asymptotically optimal performance. Nevertheless, in some cases this assumption can be removed. For more details, the reader is referred to [4].

be the GLR statistics used to reject hypothesis H_i at stage n. Let

$$N_k^{(i)} = \inf\left\{ n \ : L_k^{(i),GLR}(n) \geq B_k^{(i)} \right\}, \tag{15}$$

be the stopping rule used to reject hypothesis H_i. $B_k^{(i)}$ is the boundary value.
For each component k, the decision maker stops the sampling when $N_k = \min\left\{N_k^{(0)}, N_k^{(1)}\right\}$. If $N_k = N_k^{(0)}$, component k is declared as abnormal (i.e., H_0 is rejected). If $N_k = N_k^{(1)}$, component k is declared as normal (i.e., H_0 is accepted).

The SGLRT was first studied by Schwartz [2] for a one-parameter exponential family, who assigned a cost of c for each observation and a loss function for wrong decisions. It was shown that setting $B_k^{(i)} = \log(c^{-1})$ asymptotically minimizes the Bayes risk as c approaches zero. A refinement was studied by Lai [4], [6], who set a time-varying boundary value $B_k^{(i)} \sim \log((nc)^{-1})$. Lai showed that for a multivariate exponential family this scheme asymptotically minimizes both the Bayes risk and the expected sample size subject to error constraints as c approaches zero [6].

2) Sequential Adaptive Likelihood Ratio Test (SALRT): We refer to sequential tests that use the Adaptive Likelihood Ratio (ALR) statistics as the SALRT.
For $i = 0, 1$, let

$$L_k^{(i),ALR}(n) = \log \frac{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k(r-1))}{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k^{(i)}(n))} \tag{16}$$

be the ALR statistics used to reject hypothesis H_i at stage n. Let

$$N_k^{(i)} = \inf\left\{ n \ : L_k^{(i),ALR}(n) \geq B_k^{(i)} \right\}, \tag{17}$$

be the stopping rule used to reject hypothesis H_i, where $B_k^{(i)}$ is the boundary value.
For each component k, the decision maker stops the sampling when $N_k = \min\left\{N_k^{(0)}, N_k^{(1)}\right\}$. If $N_k = N_k^{(0)}$, component k is declared as abnormal. If $N_k = N_k^{(1)}$, component k is declared as normal.

The SALRT was first introduced by Robbins and Siegmund [3] to design power-one sequential tests. Pavlov used it to design asymptotically (as the error probability approaches zero) optimal (in terms of minimizing the expected sample size subject to error constraints) tests for composite hypothesis testing of the multivariate exponential family [5]. Tartakovsky established asymptotically optimal performance for a more general multivariate family of distributions [7].

The advantage of using the SALRT is that setting $B_k^{(0)} = \log \frac{1}{\alpha_k}$, $B_k^{(1)} = \log \frac{1}{\beta_k}$ satisfies the error probability constraints in (3). However, such a simple setting cannot be applied to the SGLRT. Thus, implementing the SALRT is much simpler than implementing the SGLRT. The disadvantage of using the SALRT is that poor early estimates (for small number of observations) can never be revised even though one has a large number of observations.

B. Asymptotically Optimal Index Policies

Under the composite hypotheses case, one should modify step 3 in Algorithms $1, 2$, given in Tables I, II by performing the SGLRT or SALRT instead of the SPRT. We refer to the modified algorithms as Algorithms $3, 4$, respectively. In the following theorems, we show that Algorithms $3, 4$ are asymptotically optimal in terms of minimizing the objective function subject to the error constraints (2) as the error probabilities approach zero[2]. When deriving asymptotics we assume that $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k such that the asymptotic optimality property in terms of minimizing the expected sample size subject to the error constraints holds for each single process for both SGLRT and SALRT, as discussed in Section V-A.

Theorem 2: Consider the independent model under the composite hypotheses case. Let $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ be the optimal solution to (2). Let $(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})$ be the solution achieved by Algorithm 3. Then, as $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k, we obtain:

$$\begin{aligned} &\boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} c_k\tau_k|(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})\right\} \\ &\qquad \sim \boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} c_k\tau_k|(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)\right\} \end{aligned} \tag{18}$$

Proof: See Appendix VIII-C. ∎

Theorem 3: Consider the exclusive model under the composite hypotheses case. Let $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ be the optimal solution to (2). Let $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ be the solution achieved by Algorithm 4. Then, as $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k, we obtain:

$$\begin{aligned} &\boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} c_k\tau_k|(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})\right\} \\ &\qquad \sim \boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} c_k\tau_k|(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)\right\} \end{aligned} \tag{19}$$

Proof: See Appendix VIII-D. ∎

C. Computing the Index

Arranging the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}(N_k)$ or $\pi_k(t_1)c_k/\mathbf{E}(N_k|H_0)$ requires one to compute the expected sample size $\mathbf{E}(N_k|H_i)$ for all $k = 1, 2, ..., K$. In general, it is difficult to obtain a closed-form expression for the exact value of $\mathbf{E}(N_k|H_i)$. However, we can use the asymptotic property of the tests to obtain a closed-form approximation to $\mathbf{E}(N_k|H_i)$, which approaches the exact expected sample size as the error probability approaches zero.

For every $i = 0, 1$, let

$$D_k(\boldsymbol{\theta}_k||\boldsymbol{\lambda}) = \mathbf{E}_{\boldsymbol{\theta}_k}\left(\log\frac{f(y_k(1)|\boldsymbol{\theta}_k)}{f(y_k(1)|\boldsymbol{\lambda})}\right) \tag{20}$$

[2]As shown in the proof of Theorems 2, 3, the index policies are still optimal in terms of testing order. The asymptotic optimality is due to the performance of the sequential test under the composite hypothesis case.

be the KL divergence between the real value of $\boldsymbol{\theta}_k$ and $\boldsymbol{\lambda}$, where the expectation is taken with respect to $f(y|\boldsymbol{\theta}_k)$, and let

$$D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(i)}) = \inf_{\boldsymbol{\lambda}\in\Theta_k^{(i)}} D_k(\boldsymbol{\theta}_k||\boldsymbol{\lambda}) \ . \tag{21}$$

Let $P^{(i)}(\boldsymbol{\theta}_k)$ be a prior distribution on $\boldsymbol{\theta}_k$ under hypothesis H_i at component k. Then, as $P_k^{FA} \to 0, P_k^{MD} \to 0$, the expected sample size is given by:

$$\begin{aligned} \mathbf{E}(N_k|H_0) &\sim \int_{\boldsymbol{\theta}_k\in\Theta_k^{(0)}} \frac{\log B_k^{(1)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(1)})} dP^{(0)}(\boldsymbol{\theta}_k) \ , \\ \mathbf{E}(N_k|H_1) &\sim \int_{\boldsymbol{\theta}_k\in\Theta_k^{(1)}\cup I_k^{(1)}} \frac{\log B_k^{(0)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(0)})} dP^{(1)}(\boldsymbol{\theta}_k) \\ &\quad + \int_{\boldsymbol{\theta}_k\in I_k^{(0)}} \frac{\log B_k^{(1)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(1)})} dP^{(1)}(\boldsymbol{\theta}_k) \ , \end{aligned} \tag{22}$$

where $I_k^{(0)}, I_k^{(1)}$ are disjoint subsets of I_k and $I_k = I_k^{(0)} \cup I_k^{(1)}$. For all $\boldsymbol{\theta}_k \in I_k^{(i)}$ we have $\frac{\log B_k^{(j)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(j)})} \leq \frac{\log B_k^{(i)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(i)})}$ for $i, j = 0, 1$.

The expected sample size required to make a decision regarding the state of component k is given by:

$$\mathbf{E}(N_k) = \pi_k\mathbf{E}(N_k|H_1) + (1-\pi_k)\mathbf{E}(N_k|H_0) \ , \tag{23}$$

which can be well approximated for small error probability using (22).

Remark 5: In numerous cases, uncertainty is associated with the abnormal state solely, where the distribution under the normal state is completely known. In these cases, evaluating $\mathbf{E}(N_k)$ to implement Algorithm 3 depends on the prior distribution of $\boldsymbol{\theta}_k \in \Theta\backslash\Theta_k^{(0)}$, while evaluating $\mathbf{E}(N_k|H_0)$ to implement Algorithm 4 does not.

VI. Numerical Examples

In this section we present numerical examples to illustrate the performance of the algorithms. Consider a cyber network consisting of K components (which can be routers, paths, etc.), as discussed in section I-B. Assume that an intruder tries to launch a DoS or Reduction of Quality (RoQ) attacks by sending a large number of packets to a component. RoQ attacks inflict damage on the component, while keeping a low profile to avoid detection. RoQ attacks do not cause denial of service.

To detect such attacks, the IDS performs a traffic-based anomaly detection. It monitors the traffic at each component to decide whether a component is compromised. Roughly speaking, if the actual arrival rate is significantly higher than the arrival rate under the normal state, then the IDS should declare that the component is in an abnormal state. A similar traffic-based detection technique was proposed in [29] for a different model, considering a single process without switching to other components. For each component k, we assume that packets arrive according to a Poisson process with rate $\theta^{(k)}$. When component k is tested, the IDS collects an observation $y_k(n) \in \mathbb{N}_0$ every time unit, which represents

the number of packets that arrived in the interval $(n-1,n)$. Assume that the IDS considers component k as normal if $\theta_k \leq \theta_k^{(0)}$, and tests $\theta_k \leq \theta_k^{(0)}$ against $\theta_k \geq \theta_k^{(1)}$ (i.e., $I_k = \{\theta_k | \theta_k^{(0)} < \theta_k < \theta_k^{(1)}\}$ is the indifference region). We set $c_k = \theta_k^{(0)}$. As discussed in Section I-B, under this setting the optimization problem minimizes the maximal damage to the network in terms of packet-loss.

A. Detection Under Simple Hypotheses

We consider the case where the observations follow Poisson distributions $y_k(n) \sim \text{Poi}(\theta_k^{(0)})$ or $y_k(n) \sim \text{Poi}(\theta_k^{(1)})$ depending on wether component k is healthy or abnormal, respectively, where $\theta_k^{(0)}, \theta_k^{(1)}$ are known to the IDS. To implement Algorithms $1, 2$ (which are optimal in this scenario for the independent and exclusive models, respectively), we need to compute the LR between the hypotheses, defined in (4), and the expected sample sizes under the hypotheses, which can be well approximated by (11). Let $\Lambda_k(n) = \log L_k(n)$ be the Log-Likelihood Ratio (LLR) between the two hypotheses of component k at stage n, where $L_k(n)$ is defined in (4). After algebraic manipulations, it can be verified that the LLR is given by:

$$\Lambda_k(n) = -n\left(\theta_k^{(1)} - \theta_k^{(0)}\right) + \log\left(\theta_k^{(1)}/\theta_k^{(0)}\right) \sum_{i=1}^{n} y_k(i) \ . \quad (24)$$

It can be verified that the KL divergence between the hypotheses H_i and H_j, defined in (10), is given by:

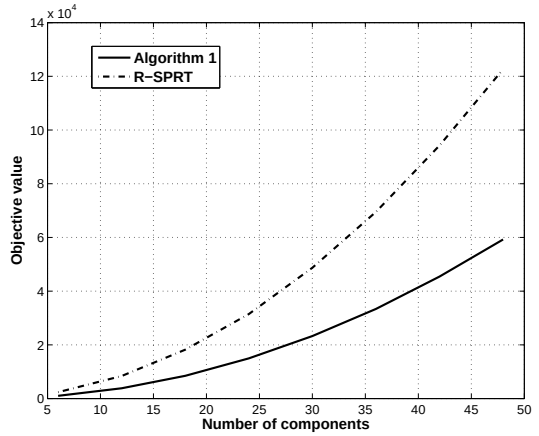
$$D_k(i||j) = \theta_k^{(j)} - \theta_k^{(i)} + \theta_k^{(i)} \log\left(\theta_k^{(i)}/\theta_k^{(j)}\right) \ . \quad (25)$$

Substituting (25) in (11) yields the required approximation to the expected sample size.

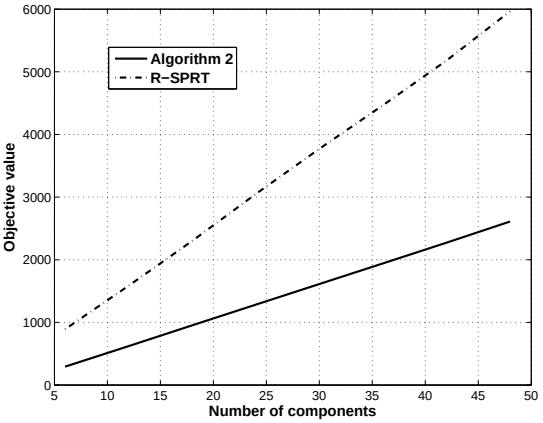
Next, we provide numerical examples to illustrate the performance of the algorithms. We compared three schemes: a Random selection SPRT (R-SPRT), where a series of SPRTs are performed until all the components are tested in a random order (which is optimal for the problem of minimizing the detection delay over independent processes [11]), and the proposed Algorithms $1, 2$, which are optimal under the independent and exclusive models, respectively.

Let $\Delta_K = (100-10)/(K-1)$. We set $c_k = \theta_k^{(0)} = 10 + (k-1)\Delta_K$ (i.e., the costs are equally spaced in the interval $[10, 100]$) and $\theta_k^{(1)} = 1.5 \cdot \theta_k^{(0)}$. The error constraints were set to $P_k^{FA} = 10^{-2}, P_k^{MD} = 10^{-6}$ for all k. For the independent and exclusive models, we set $\pi_k = 0.8$ and $\pi_k = 1/K$ for all k, respectively. The performance of Algorithms 1 and 2 are presented in Fig. 1(a) and 1(b) under the independent and exclusive models, respectively, and compared to the R-SPRT. It can be seen that the proposed Algorithms save roughly 50% of the objective value as compared to the R-SPRT under both the independent and exclusive model scenarios.

Next, we simulate the independent model when 2 components are observed at a time and the total number of components is $K = 6$. Note that in this case Algorithm 1 may not be optimal. We use an exhaustive search as a bench mark to demonstrate the performance of Algorithm 1 in this scenario. The exhaustive search is done by performing a sequence of



(a) An independent model scenario.



(b) An exclusive model scenario.

Fig. 1. Objective value as a function of the number of components under the independent and exclusive models.

K SPRTs among all the possible testing orders. Then, the minimal objective value is chosen as a bench mark. We set the maximal cost to $c_{max} = 100$ and the costs are equally spaced in the interval $[c_{min}, 100]$. The error constraints were set to $P_k^{FA} = P_k^{MD} = 10^{-2}$ for all k. The performance gain of the exhaustive search scheme over Algorithm 1 as a function of c_{min} are presented in Fig. 2. It can be seen that Algorithm 1 almost achieves the performance of the exhaustive search scheme in this scenario for all c_{min}. For small c_{min} both algorithms perform the same, since the difference between the indices increases. The exhaustive search outperforms Algorithm 1 for $c_{min} > 97$, but the gain remains very small.

B. Detection Under Uncertainty

We consider the case of composite hypotheses, where there is uncertainty in the distribution parameters, as discussed in Section V. To implement the asymptotically optimal Algorithms $3, 4$, we need to compute the GLR or ALR statistics, defined in (14), (16) and the expected sample sizes under the hypotheses, which can be well approximated by (22). The MLEs of the parameters over the parameter spaces Θ_k, $\Theta_k^{(i)}$ are given by the sample mean and the boundary of the alternative parameter space, respectively. As a result,

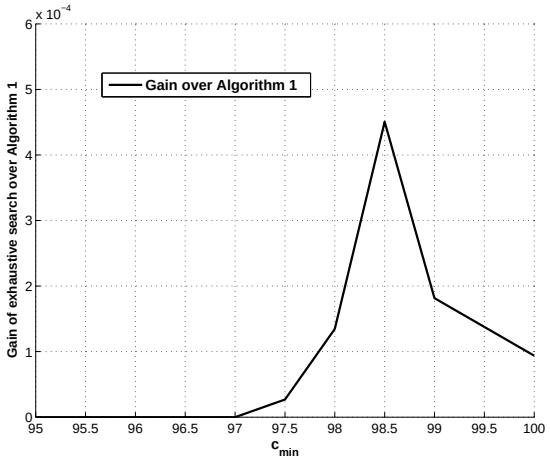


Fig. 2. Performance gain of an exhaustive search over Algorithm 1 as a function of c_{min} under the independent model.

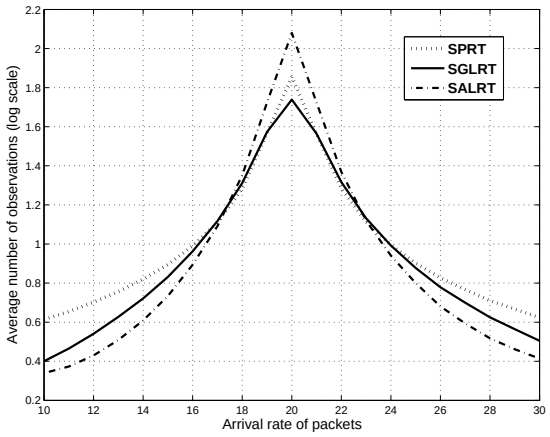


Fig. 3. Average number of observations as a function of the arrival rate of packets (denoted by θ).

substituting: $\hat{\theta}_k(n) = \frac{1}{n}\sum_{i=1}^{n} y_k(i)$, $\hat{\theta}_k^{(i)}(n) = \theta_k^{(i)}$, in (14), (16) yields the GLR and ALR statistics, respectively. The KL divergence between the real value of θ_k and the parameter space $\Theta_k^{(i)}$ is given by:

$$D_k^*(\theta_k||\Theta_k^{(i)}) = \theta_k^{(i)} - \theta_k + \theta_k \log\left(\theta_k/\theta_k^{(i)}\right) . \qquad (26)$$

Substituting (26) in (22) yields the approximate expected sample size.

Next, we provide numerical examples to illustrate the performance of the algorithms under uncertainty. We simulated a network with homogenous components (i.e., any selection rule is optimal). We compared three schemes: R-SPRT, and Algorithms 3 or 4 (which achieve the same performance in this case) using the SALRT and the SGLRT, discussed in section V-A. We set $\theta_k^{(0)} = 19$, $\theta_k^{(1)} = 21$. Under uncertainty, the IDS considers component k as normal if $\theta_k \leq \theta_k^{(0)}$, and tests $\theta_k \leq \theta_k^{(0)}$ against $\theta_k \geq \theta_k^{(1)}$ (i.e., $I_k = \{\theta_k | 19 < \theta_k < 21\}$ is the indifference region). To implement the SGLRT, we set the cost per observation $c = 10^{-3}$. According to the assigned cost, we obtained the following error probability constraints for all k: $P_k^{FA} \leq 0.026$ for all $\theta^{(k)} \leq 19$ and $P_k^{MD} \leq 0.03$ for all $\theta^{(k)} \geq 21$. We do not restrict the detector's performance for $19 < \theta^{(k)} < 21$ (Note that narrowing the indifference region has the price of increasing the required sample size). In Fig. 3 we show the average number of observations (in a log scale) required for the anomaly detection as a function of $\theta^{(k)}$. As expected, for $\theta_k = 19$ and $\theta_k = 21$ the R-SPRT requires lower sample size as compared to the proposed schemes. On the other hand, it can be seen that for most values of θ the SGLRT and the SALRT require lower sample size as compared to the R-SPRT. The SALRT performs the worst for $18 < \theta_k < 22$, and performs the best for $\theta_k \notin (18, 22)$, roughly. The SGLRT obtains the best average performance. It can be seen that for large values of θ_k the anomaly is detected very quickly, since the distance between the hypotheses increases. This result confirms that DoS attacks are much easier to detect than RoQ attacks.

VII. Conclusion

The problem of anomaly localization in a resource-constrained cyber system was investigated. Due to resource constraints, only one component can be probed at a time. The observations are realizations drawn from two different distributions depending on whether the component is normal or anomalous. An abnormal component incurs a cost per unit time until it is tested and identified. The problem was formulated as a constrained optimization problem. The objective is to minimize the total expected cost subject to error probability constraints. We considered two different anomaly models: the independent model in which each component can be abnormal independent of other components, and the exclusive model in which there is one and only one abnormal component. For the simple hypotheses case, we derived optimal algorithms for both independent and exclusive models. For the composite hypotheses case, we derived asymptotically (as the error probability approaches zero) optimal algorithms for both independent and exclusive models. These optimal algorithms have low-complexity.

The algorithms that have been developed in this paper can be applied to other models of anomaly detection as well. We can modify the proposed algorithms to any detection scheme that performs a series of tests until all the components are tested. The required modification is in step 3 of the algorithms, where the SPRT/SALRT/SGLRT are replaced by any given test. Such modified algorithms minimize the objective function among all the algorithms that perform the given test.

VIII. Appendix

A. Proof of Theorem 1 Under The Exclusive Model

Let $\mathbf{E}'(N_k|H_i,\mathrm{t})$ be the expected sample size achieved by a stopping rule and a decision rule $(\tau_k'(t), \delta_k'(t))$, depending on the time that component k is tested (i.e., $(\tau_k'(t), \delta_k'(t))$ depend on the selection rule), such that error constraints are satisfied. Let $\mathbf{E}^{A2}(N_k|H_i)$ be the expected sample size achieved by the SPRT's stopping rule and decision rule $(\tau_k^{A2}, \delta_k^{A2})$, independent of the time that component k is tested (i.e., $(\tau_k^{A2}, \delta_k^{A2})$ are independent of the selection rule), such that error constraints

are satisfied. Clearly, $\mathbf{E}^{A2}(N_k|H_i) \leq \mathbf{E}'(N_k|H_i,t)$ for all k,t, for $i=0,1$.

Step 1***:*** *Proving the theorem for* $K=2$*:*

Assume that

$$\frac{\pi_1(t_1)c_1}{\mathbf{E}^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^{A2}(N_2|H_0)} \ . \tag{27}$$

Consider selection rules $\boldsymbol{\phi}^{(1)}$, $\boldsymbol{\phi}^{(2)}$ that select component 1 first followed by component 2 and component 2 first followed by component 1, respectively. The expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(2)})\right\}\\ &= \left(\mathbf{E}'(N_2|H_1,t_1)\right)\pi_2(t_1)c_2\\ &\quad + \left(\mathbf{E}'(N_2|H_0,t_1)+\mathbf{E}'(N_1|H_1,t_2)\right)\pi_1(t_1)c_1.\end{aligned} \tag{28}$$

The expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(1)})\right\}\\ &= \left(\mathbf{E}'(N_1|H_1,t_1)\right)\pi_1(t_1)c_1\\ &\quad + \left(\mathbf{E}'(N_1|H_0,t_1)+\mathbf{E}'(N_2|H_1,t_2)\right)\pi_2(t_1)c_2.\end{aligned} \tag{29}$$

Note that the expected cost achieved by both selection rules can be further reduced by minimizing the expected sample sizes (such that error constraints are satisfied) independent of the selection rules, which is achieved by $(\tau_k^{A2},\delta_k^{A2})$. Therefore, an optimal solution must be $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})$ or $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(2)})$. Next, we use the interchange argument to prove the theorem for $K=2$. The expected cost achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(2)})\right\}\\ &= \left(\mathbf{E}^{A2}(N_2|H_1)\right)\pi_2(t_1)c_2\\ &\quad + \left(\mathbf{E}^{A2}(N_2|H_0)+\mathbf{E}^{A2}(N_1|H_1)\right)\pi_1(t_1)c_1.\end{aligned} \tag{30}$$

The expected cost achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})\right\}\\ &= \left(\mathbf{E}^{A2}(N_1|H_1)\right)\pi_1(t_1)c_1\\ &\quad + \left(\mathbf{E}^{A2}(N_1|H_0)+\mathbf{E}^{A2}(N_2|H_1)\right)\pi_2(t_1)c_2.\end{aligned} \tag{31}$$

the expected cost achieved by $\boldsymbol{\phi}^{(1)}$ is lower than that achieved by $\boldsymbol{\phi}^{(2)}$ since $\frac{\pi_1(t_1)c_1}{\mathbf{E}^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^{A2}(N_2|H_0)}$, which completes the proof for $K=2$.

Step 2***:*** *Proving the theorem by induction on the number of components* K*:*

Assume that the theorem is true for $K-1$ components (where one and only one component is abnormal). Assume that

$$\frac{\pi_1(t_1)c_1}{\mathbf{E}^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^{A2}(N_2|H_0)} \geq ... \geq \frac{\pi_K(t_1)c_K}{\mathbf{E}^{A2}(N_K|H_0)} \ . \tag{32}$$

Consider the case of K components and denote $\boldsymbol{\phi}^{(j)}$ as an optimal selection rule that selects component j first.

Step 2.1***:*** *Proving the theorem for the last* $K-1$ *components:*

Next, we show that the last $K-1$ components must be selected in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}^{A2}(N_k|H_0)$ and tested by the SPRT.

Let

$$\gamma_j(t) = \frac{1}{\pi_j(t)\frac{f_j^{(1)}(\mathbf{y}_j(N_j))}{f_j^{(0)}(\mathbf{y}_j(N_j))} + 1 - \pi_j(t)} \ . \tag{33}$$

Note that when the decision maker completes testing component j, the other components update their beliefs according to:

$$\pi_k(t_2) = \gamma_j(t_1)\pi_k(t_1) \ , \ \forall k \neq j \ . \tag{34}$$

The expected cost achieved by $\boldsymbol{\phi}^{(j)}$ given the outcome (at time t_2) by testing component j (i.e., given the observations vector $\mathbf{y}_j(N_j)$) is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j)\right\}\\ &= \pi_j(t_2)c_jN_j + (1-\pi_j(t_2))\times\\ &\quad \mathbf{E}\left\{\sum_{k=1,k\neq j}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j), j\in\mathcal{H}_0\right\} \ .\end{aligned} \tag{35}$$

Let

$$\widetilde{\tau}_k = \tau_k - N_j \quad \forall k \neq j \tag{36}$$

be the modified stopping time, defined as the stopping time from $t=N_j+1$ until testing of component k is completed. Thus, we can rewrite (35) as:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j)\right\}\\ &= \sum_{k=1}^{K}\pi_k(t_2)c_kN_j + (1-\pi_j(t_2))\times\\ &\quad \mathbf{E}\left\{\sum_{k=1,k\neq j}^{K} c_k\widetilde{\tau}_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j), j\in\mathcal{H}_0\right\} \ .\end{aligned} \tag{37}$$

The term $\sum_{k=1}^{K}\pi_k(t_2)c_kN_j$ in (37) follows since,

$$\begin{aligned}&\Pr\left(k\in\mathcal{H}_1 \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),j\in\mathcal{H}_0\right)\\&=\frac{\Pr\left(k\in\mathcal{H}_1,j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),\right)}{\Pr\left(j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),\right)}\\&=\frac{\Pr\left(k\in\mathcal{H}_1 \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),\right)}{\Pr\left(j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),\right)}=\frac{\pi_k(t_2)}{1-\pi_j(t_2)}\triangleq\widetilde{\pi}_k(t_2)\ .\end{aligned}\tag{38}$$

Minimizing

$$\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j)\right\}\tag{39}$$

at time t_2, requires one to minimize

$$\mathbf{E}\left\{\sum_{k=1,k\neq j}^{K}c_k\widetilde{\tau}_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)},\mathbf{y}_j(N_j),j\in\mathcal{H}_0\right\}\tag{40}$$

in (37).

Note that (40) is the cost for $K-1$ components (where one and only one component is abnormal) starting at time $t=t_2=N_j+1$, with prior probability $\widetilde{\pi}_k(t_2)=\frac{\pi_k(t_2)}{1-\pi_j(t_2)}$ for component $k\neq j$ being abnormal. By the induction hypothesis, for any optimal selection rule $\boldsymbol{\phi}^{(j)}$ that selects component j first, arranging the last $K-1$ components in decreasing order of $\widetilde{\pi}_k(t_2)c_k/\mathbf{E}^{A2}(N_k|H_0)$ (and testing them by the SPRT) minimizes (40).

Since

$$\widetilde{\pi}_k(t_2)=\frac{\gamma_j(t_1)}{1-\pi_j(t_2)}\pi_k(t_1)\ \ \forall k\neq j,\tag{41}$$

then

$$\begin{aligned}&\frac{\widetilde{\pi}_1(t_2)c_1}{\mathbf{E}^{A2}(N_1|H_0)}\geq\frac{\widetilde{\pi}_2(t_2)c_2}{\mathbf{E}^{A2}(N_2|H_0)}\geq\cdots\geq\frac{\widetilde{\pi}_{j-1}(t_2)c_{j-1}}{\mathbf{E}^{A2}(N_{j-1}|H_0)}\\&\geq\frac{\widetilde{\pi}_{j+1}(t_2)c_{j+1}}{\mathbf{E}^{A2}(N_{j+1}|H_0)}\geq\cdots\geq\frac{\widetilde{\pi}_K(t_2)c_K}{\mathbf{E}^{A2}(N_K|H_0)}.\end{aligned}\tag{42}$$

Thus, the last $K-1$ components must be selected in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}^{A2}(N_k|H_0)$ and tested by the SPRT.

Step 2.2***: Proving the theorem for all the*** K ***components:***

Finally, we show that component 1 (i.e., the component with the highest index) must be selected first. The expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})\right\}\\&=\pi_j(t_1)c_j\left(\mathbf{E}'(N_j|H_1,t_1)\right)+\sum_{k=1,k\neq j}^{K}[\pi_k(t_1)c_k\times\\&\left(\mathbf{E}'\left(N_j|H_0,t_1\right)+\left(\sum_{i=1,i\neq j}^{k-1}\mathbf{E}^{A2}\left(N_i|H_0\right)\right)\right.\\&\left.+\mathbf{E}^{A2}\left(N_k|H_1\right)\right)\Big]\ .\end{aligned}\tag{43}$$

First, note that the expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})$ can be further reduced for all j by minimizing the expected sample size $\mathbf{E}'(N_j|H_i,t_1)$ for $i=0,1$, which is achieved by $(\tau_j^{A2},\delta_j^{A2})$. Therefore, an optimal solution must be $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j)})$ for an optimal selection rule $\boldsymbol{\phi}^{(j)}$. Thus, in the following we consider solutions of the form $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi})$.

Next, by contradiction, consider an optimal selection rule $\boldsymbol{\phi}^{(j\neq1)}$ that selects component $j\neq1$ first. Therefore, $\boldsymbol{\phi}^{(j\neq1)}$ selects the components in the following order:

$$j,1,2,...,j-1,j+1,...,K.$$

As a result, the expected cost achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})\right\}\\&=\pi_j(t_1)c_j\left(\mathbf{E}^{A2}(N_j|H_1)\right)\\&\quad+\pi_1(t_1)c_1\left[\mathbf{E}^{A2}\left(N_j|H_0\right)+\mathbf{E}^{A2}\left(N_1|H_1\right)\right]\\&\quad+\sum_{k=2,k\neq j}^{K}[\pi_k(t_1)c_k\times\\&\quad\left(\mathbf{E}^{A2}\left(N_j|H_0\right)+\left(\sum_{i=1,i\neq j}^{k-1}\mathbf{E}^{A2}\left(N_i|H_0\right)\right)\right.\\&\quad\left.+\mathbf{E}^{A2}\left(N_k|H_1\right)\right)\Big]\ .\end{aligned}\tag{44}$$

We use the interchange argument to prove the theorem. Consider a selection rule $\boldsymbol{\phi}^{(1)}$ that selects component 1 first followed by components $j,2,3,j-1,j+1,...,K$. Similar to (44), the expected cost achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})\right\}\\&=\pi_1(t_1)c_1\left(\mathbf{E}^{A2}(N_1|H_1)\right)\\&\quad+\pi_j(t_1)c_j\left[\mathbf{E}^{A2}\left(N_1|H_0\right)+\mathbf{E}^{A2}\left(N_j|H_1\right)\right]\\&\quad+\sum_{k=2,k\neq j}^{K}[\pi_k(t_1)c_k\times\\&\quad\left(\mathbf{E}^{A2}\left(N_j|H_0\right)+\left(\sum_{i=1,i\neq j}^{k-1}\mathbf{E}^{A2}\left(N_i|H_0\right)\right)\right.\\&\quad\left.+\mathbf{E}^{A2}\left(N_k|H_1\right)\right)\Big]\ .\end{aligned}\tag{45}$$

By comparing (44) and (45), it can be verified that:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})\right\}\\&\qquad\leq\mathbf{E}\left\{\sum_{k=1}^{K}c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})\right\}\end{aligned}$$

since $\pi_1(t_1)c_1/\mathbf{E}^{A2}(N_1|H_0) \geq \pi_j(t_1)c_j/\mathbf{E}^{A2}(N_j|H_0)$.
The expected cost can be reduced by selecting component 1 first followed by component j, which contradicts the optimality of $\boldsymbol{\phi}^{(j\neq 1)}$. Hence, at time t_1 selecting component 1 minimizes the expected cost. We have already proved that selecting the last $K-1$ components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}^{A2}(N_k|H_0)$ minimizes the objective function, which completes the proof. ■

B. Proof of Theorem 1 Under The Independent Model

Let $\mathbf{E}'(N_k|H_i,\mathrm{t})$ be the expected sample size achieved by a stopping rule and a decision rule $(\tau_k'(t),\delta_k'(t))$, depending on the time that component k is tested (i.e., $(\tau_k'(t),\delta_k'(t))$ depend on the selection rule), such that error constraints are satisfied. Let $\mathbf{E}^{A1}(N_k|H_i)$ be the expected sample size achieved by the SPRT's stopping rule and decision rule $(\tau_k^{A1},\delta_k^{A1})$, independent of the time that component k is tested (i.e., $(\tau_k^{A1},\delta_k^{A1})$ are independent of the selection rule), such that error constraints are satisfied. Clearly, $\mathbf{E}^{A1}(N_k|H_i) \leq \mathbf{E}'(N_k|H_i,t)$ for all k,t, for $i=0,1$ and are achieved by Algorithm 1.

First, consider the case where $K=2$. Assume that

$$\frac{\pi_1(t_1)c_1}{\mathbf{E}^{A1}(N_1)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^{A1}(N_2)} .$$

Consider selection rules $\boldsymbol{\phi}^{(1)}$, $\boldsymbol{\phi}^{(2)}$ that select component 1 first followed by component 2 and component 2 first followed by component 1, respectively. The expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(2)})\right\}\\ &\quad=\left(\mathbf{E}'(N_2|H_1,t_1)\right)\pi_2(t_1)c_2\\ &\qquad+\left(\mathbf{E}'(N_2|t_1)+\mathbf{E}'(N_1|H_1,t_2)\right)\pi_1(t_1)c_1.\end{aligned} \tag{46}$$

The expected cost achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(1)})\right\}\\ &\quad=\left(\mathbf{E}'(N_1|H_1,t_1)\right)\pi_1(t_1)c_1\\ &\qquad+\left(\mathbf{E}'(N_1|t_1)+\mathbf{E}'(N_2|H_1,t_2)\right)\pi_2(t_1)c_2.\end{aligned} \tag{47}$$

Note that the expected cost achieved by both selection rules can be further reduced by minimizing the expected sample sizes (such that error constraints are satisfied) independent of the selection rules, which is achieved by $(\tau_k^{A1},\delta_k^{A1})$. Therefore, an optimal solution must be $(\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(1)})$ or $(\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(2)})$. Next, we use the interchange argument to prove the theorem for $K=2$. The expected cost achieved by $(\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(2)})\right\}\\ &\quad=\left(\mathbf{E}^{A1}(N_2|H_1)\right)\pi_2(t_1)c_2\\ &\qquad+\left(\mathbf{E}^{A1}(N_2)+\mathbf{E}^{A1}(N_1|H_1)\right)\pi_1(t_1)c_1.\end{aligned} \tag{48}$$

The expected cost achieved by $(\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1},\boldsymbol{\delta}^{A1},\boldsymbol{\phi}^{(1)})\right\}\\ &\quad=\left(\mathbf{E}^{A1}(N_1|H_1)\right)\pi_1(t_1)c_1\\ &\qquad+\left(\mathbf{E}^{A1}(N_1)+\mathbf{E}^{A1}(N_2|H_1)\right)\pi_2(t_1)c_2.\end{aligned} \tag{49}$$

The expected cost achieved by $\boldsymbol{\phi}^{(1)}$ is lower than that achieved by $\boldsymbol{\phi}^{(2)}$ since $\frac{\pi_1(t_1)c_1}{\mathbf{E}^{A1}(N_1)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^{A1}(N_2)}$, which completes the proof for $K=2$.
The rest of the proof follows by induction on the number of components, as was done under the exclusive model. ■

C. Proof of Theorem 2

For every k, let $\mathbf{E}^*(N_k|H_i)$ be the minimal expected sample size that can be achieved by any sequential test, such that error constraints are satisfied. Let $\mathbf{E}^{A3}(N_k|H_i)$ be the expected sample size achieved by Algorithm 3, such that error constraints are satisfied. Clearly, $\mathbf{E}^*(N_k|H_i) \leq \mathbf{E}^{A3}(N_k|H_i)$ for all k, for $i=0,1$.
Assume that

$$\frac{\pi_1(t_1)c_1}{\mathbf{E}^*(N_1)} \geq \frac{\pi_2(t_1)c_2}{\mathbf{E}^*(N_2)} \geq ... \geq \frac{\pi_K(t_1)c_K}{\mathbf{E}^*(N_K)}. \tag{50}$$

Similar to the proof of Theorem 1, it can be verified that the optimal solution to (2) is to select the components in the following order: $1,2,...,K$, where the components are tested by a sequential test that achieves expected sample size $\mathbf{E}^*(N_k|H_i)$ for all k, for $i=0,1$. Therefore, the expected cost achieved by $(\boldsymbol{\tau}^*,\boldsymbol{\delta}^*,\boldsymbol{\phi}^*)$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^*,\boldsymbol{\delta}^*,\boldsymbol{\phi}^*)\right\}\\ &=\sum_{k=1}^{K}\pi_k(t_1)c_k\left[\left(\sum_{i=1}^{k-1}\mathbf{E}^*\left(N_i\right)\right)+\mathbf{E}^*\left(N_k|H_1\right)\right].\end{aligned} \tag{51}$$

By the asymptotic optimality property of the SALRT/SGLRT for a single process (used in Algorithm 3), it follows that $\mathbf{E}^{A3}(N_k|H_i) \sim \mathbf{E}^*(N_k|H_i)$ for all k, for $i=0,1$ as $P_k^{FA}\to 0, P_k^{MD}\to 0$. As a result, for sufficiently small error probabilities, the solution $(\boldsymbol{\tau}^{A3},\boldsymbol{\delta}^{A3},\boldsymbol{\phi}^{A3})$ is to select the components in the following order: $1,2,...,K$, where the components are tested by an asymptotically optimal sequential test that achieves expected sample size $\mathbf{E}^{A3}(N_k|H_i)$ for all k, for $i=0,1$. Therefore, the expected cost achieved by $(\boldsymbol{\tau}^{A3},\boldsymbol{\delta}^{A3},\boldsymbol{\phi}^{A3})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} c_k\tau_k\mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A3},\boldsymbol{\delta}^{A3},\boldsymbol{\phi}^{A3})\right\}\\ &=\sum_{k=1}^{K}\pi_k(t_1)c_k\left[\left(\sum_{i=1}^{k-1}\mathbf{E}^{A3}\left(N_i\right)\right)+\mathbf{E}^{A3}\left(N_k|H_1\right)\right].\end{aligned} \tag{52}$$

Since $\mathbf{E}^{A3}(N_k|H_i) \sim \mathbf{E}^*(N_k|H_i)$ for $i=0,1$ as $P_k^{FA}\to 0, P_k^{MD}\to 0$ for all k, the theorem follows. ■

D. Proof of Theorem 3

The structure of the proof is similar to the proof of Theorem 2. Hence, we provide a sketch of the proof, using notation similar to that used in the proof of Theorem 2. Similar to the proof of Theorem 1, it can be verified that the optimal solution to (2) is to select the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}^*(N_k|H_0)$, where the components are tested by a sequential test that achieves expected sample size $\mathbf{E}^*(N_k|H_i)$ for all k, for $i = 0, 1$. By the asymptotic optimality property for a single process of the SALRT/SGLRT (used in Algorithm 4), it follows that $\mathbf{E}^{A4}(N_k|H_i) \sim \mathbf{E}^*(N_k|H_i)$ for all k, for $i = 0, 1$ as $P_k^{FA} \to 0, P_k^{MD} \to 0$. As a result, for sufficiently small error probabilities, the solution $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ is to select the components in decreasing order of $\pi_k(t_1)c_k/\mathbf{E}^*(N_k|H_0)$, where the components are tested by an asymptotically optimal sequential test that achieves expected sample size $\mathbf{E}^{A4}(N_k|H_i)$ for all k, for $i = 0, 1$. Similar to the proof of Theorem 2, comparing the objective functions achieved by $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ and $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ proves the theorem. ■

REFERENCES

[1] A. Wald, "Sequential analysis," *New York: Wiley*, 1947.

[2] G. Schwarz, "Asymptotic shapes of Bayes sequential testing regions," *The Annals of mathematical statistics*, pp. 224–236, 1962.

[3] H. Robbins and D. Siegmund, "The expected sample size of some tests of power one," *The Annals of Statistics*, pp. 415–436, 1974.

[4] T. L. Lai, "Nearly optimal sequential tests of composite hypotheses," *The Annals of Statistics*, pp. 856–886, 1988.

[5] I. V. Pavlov, "Sequential procedure of testing composite hypotheses with applications to the Kiefer-Weiss problem," *Theory of Probability and Its Applications*, vol. 35, no. 2, pp. 280–292, 1990.

[6] T. L. Lai and L. M. Zhang, "Nearly optimal generalized sequential likelihood ratio tests in multivariate exponential families," *Lecture Notes-Monograph Series*, pp. 331–346, 1994.

[7] A. G. Tartakovsky, "An efficient adaptive sequential procedure for detecting targets," in *IEEE Aerospace Conference Proceedings, 2002*, vol. 4, pp. 1581–1596, 2002.

[8] V. Draglin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests - part i: Asymptotic optimality," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2448–2461, 1999.

[9] H. Li, "Restless watchdog: Selective quickest spectrum sensing in multichannel cognitive radio systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[10] Q. Zhao and J. Ye, "Quickest detection in multiple on–off processes," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 5994–6006, 2010.

[11] R. Caromi, Y. Xin, and L. Lai, "Fast multiband spectrum scanning for cognitive radio systems," *IEEE Transaction on Communications*, vol. 61, no. 1, pp. 63–75, 2013.

[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] M. L. Malloy, G. Tang, and R. D. Nowak, "Quickest search for a rare distribution," *IEEE Annual Conference on Information Sciences and Systems*, pp. 1–6, 2012.

[14] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4462–4481, 2013.

[15] K. S. Zigangirov, "On a problem in optimal scanning," *Theory of Probability and Its Applications*, vol. 11, no. 2, pp. 294–298, 1966.

[16] E. Klimko and J. Yackel, "Optimal search strategies for Wiener processes," *Stochastic Processes and their Applications*, vol. 3, no. 1, pp. 19–33, 1975.

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

[18] L. D. Stone and J. A. Stanshine, "Optimal search using uninterrupted contact investigation," *SIAM Journal on Applied Mathematics*, vol. 20, no. 2, pp. 241–263, 1971.

[19] K. P. Tognetti, "An optimal strategy for a whereabouts search," *Operations Research*, vol. 16, no. 1, pp. 209–211, 1968.

[20] J. B. Kadane, "Optimal whereabouts search," *Operations Research*, vol. 19, no. 4, pp. 894–904, 1971.

[21] D. A. Castanon, "Optimal search strategies in dynamic hypothesis testing," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 7, pp. 1130–1138, 1995.

[22] Y. Zhai and Q. Zhao, "Dynamic search under false alarms," *to appear in the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2013.

[23] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 265–272, 1953.

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

[25] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *The Annals of Mathematical Statistics*, pp. 404–419, 1962.

[26] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *to appear in Annals of Statistics*.

[27] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *to appear in the IEEE Transactions on Automatic Control*, 2013.

[28] W. E. Smith, "Various optimizers for single-stage production," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 59–66, 1956.

[29] I. Onat and A. Miri, "An intrusion detection system for wireless sensor networks," in *IEEE International Conference on Wireless And Mobile Computing, Networking And Communications*, vol. 3, pp. 253–259, 2005.