This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

Sequential anomaly detection under sampling constraints

Aristomenis Tsopelakos and Georgios Fellouris, *Member, IEEE*

***Abstract*—The problem of sequential anomaly detection is considered, where multiple data sources are monitored in real time and the goal is to identify the "anomalous" ones among them, when it is not possible to sample all sources at all times. A detection scheme in this context requires specifying not only when to stop sampling and which sources to identify as anomalous upon stopping, but also which sources to sample at each time instance until stopping. A novel formulation for this problem is proposed, in which the number of anomalous sources is not necessarily known in advance and the number of sampled sources per time instance is not necessarily fixed. Instead, an arbitrary lower bound and an arbitrary upper bound are assumed on the number of anomalous sources, and the fraction of the expected number of samples over the expected time until stopping is required to not exceed an arbitrary, user-specified level. In addition to this sampling constraint, the probabilities of at least one false alarm and at least one missed detection are controlled below user-specified tolerance levels. A general criterion is established for a policy to achieve the minimum expected time until stopping to a first-order asymptotic approximation as the two familywise error rates go to zero. Moreover, the asymptotic optimality is established of a family of policies that sample each source at each time instance with a probability that depends on past observations only through the current estimate of the subset of anomalous sources. This family includes, in particular, a novel policy that requires minimal computation under any setup of the problem.**

***Index Terms*—Active sensing; Anomaly detection; Asymptotic optimality; Controlled sensing; Sequential design of experiments; Sequential detection; Sequential sampling; Sequential testing.**

I. Introduction

In various engineering and scientific areas data are often collected in real time over multiple streams, and it is of interest to quickly identify those data streams, if any, that exhibit outlying behavior. In brain science, for example, it is desirable to identify groups of cells with large vibration frequency, as this is a symptom for the development of a particular malfunctioning [1]. In fraud prevention security systems in e-commerce, it is desirable to identify transition links with low transition rate, as this may be an indication that a link is tapped [2]. Such applications, among many others, motivate the study of sequential multiple testing problems where the data for the various hypotheses are generated by distinct sources, there are two hypotheses for each data source, and the goal is to identify as quickly as possible the "anomalous" sources, i.e., those in which the alternative hypothesis is correct. In certain works, e.g., [3]–[8], all sources are sampled at each time instance, whereas in others, e.g., [9]–[17], only a fixed number of sources (typically, only one). In the latter case, apart from when to stop sampling and which data sources to identify as anomalous upon stopping, one also needs to specify which sources to sample at every time instance until stopping.

The latter problem, which is often called *sequential anomaly detection* in the literature, can be viewed as a special case of the *sequential multi-hypothesis testing problem with controlled sensing (or observation control)*, where the goal is to solve a sequential multi-hypothesis testing problem while taking at each time instance an action that influences the distribution of the observations [18]–[28]. In the anomaly detection case, the action is the selection of the sources to be sampled, whereas the hypotheses correspond to the possible subsets of anomalous sources. Therefore, policies and results in the context of sequential multi-hypothesis testing with controlled sensing are applicable, in principle at least, to the sequential anomaly detection problem. Such a policy was first proposed in [18] for the case of two hypotheses, and subsequently generalized in [20], [21] to the case of an arbitrary, finite number of hypotheses. When applied to the sequential anomaly detection problem, this policy samples each subset of sources of the allowed size at each time instance with a certain probability that depends on the past observations only through the currently estimated subset of anomalous sources.

In general, the implementation of the policy in [20]

Aristomenis Tsopelakos is with the Electrical and Computer Engineering Department and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana IL 61801, USA, (e-mail: tsopela2@illinois.edu). Georgios Fellouris is with the Department of Statistics, the Electrical and Computer Engineering Department, and the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Champaign, IL 61820 USA, (e-mail: fellouri@illinois.edu) The work of the two authors were supported in part by the NSF under Grant CIF 1514245, through the University of Illinois at Urbana–Champaign. The work of the second author was also supported in part by the NSF under Grant DMS 1737962, through the University of Illinois at Urbana–Champaign.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

requires solving, for each subset of anomalous sources, a linear system where the number of equations is equal to the number of sources and the number of unknowns is equal to the number of all subsets of sources of the allowed size. Moreover, its asymptotic optimality has been established only under restrictive assumptions, such as when the following hold simultaneously: it is known *a priori* that there is only *one* anomalous source, it is possible to sample only *one* source at a time, the testing problems are identical, and the sources generate *Bernoulli* random variables under each hypothesis [21, Appendix A]. To avoid such restrictions, it has been proposed to modify the policy in [20] at an appropriate subsequence of time instances, at which each subset of sources of the allowed size is sampled with the same probability [18, Remark 7], [25]. Such a *modified* policy was shown in [25] to always be asymptotically optimal, as long as the log-likelihood ratio statistic of each observation has a finite *second* moment.

A goal of the present work is to show that the *unmodified* policy in [20] is always asymptotically optimal in the context of the above sequential anomaly detection problem, as long as the log-likelihood ratio statistic of each observation has a finite *first* moment. However, our main goal in this paper is to propose a more general framework for the problem of sequential anomaly detection with sampling constraints that (i) does not rely on the restrictive assumption that the number of anomalous sources is known in advance, (ii) allows for two distinct error constraints and captures the asymmetry between a *false alarm*, i.e., falsely identifying a source as anomalous, and a *missed detection*, i.e., failing to detect an anomalous source, and most importantly, (iii) relaxes the hard sampling constraint that the same number of sources must be sampled at each time instance, and (iv) admits an asymptotically optimal solution that is convenient to implement under any setup of the problem.

To be more specific, in this paper we assume an arbitrary, user-specified lower bound and an arbitrary, user-specified upper bound on the number of anomalous sources. This setup includes the case of no prior information, the case where the number of anomalous sources is known in advance, as well as more realistic cases of prior information, such as when there is only a non-trivial upper bound on the number of anomalous sources. Moreover, we require control of the probabilities of at least one false alarm and at least one missed detection below arbitrary, user-specified levels. Both these features are taken into account in [6] when all sources are observed at all times. Thus, the present paper can be seen as a generalization of [6] to the case that it is not possible to observe all sources at all times. However, instead of demanding that the number of sampled sources per time instance be fixed, as in [9]–[17], we only require that *the ratio of the expected number of observations over the expected time until stopping* not exceed a user-specified level. This leads to a more general formulation for sequential anomaly detection compared to those in [9]–[17], which at the same time is *not* a special case of the sequential multi-hypothesis testing problem with controlled sensing in [18]–[28]. Thus, while existing policies in the literature are applicable to the proposed setup, this is not the case for the existing universal lower bounds.

Our first main result on the proposed problem is a criterion for a policy (that employs the stopping and decision rules in [6] and satisfies the sampling constraint) to achieve the optimal expected time until stopping to a first-order asymptotic approximation as the two familywise error probabilities go to 0. Indeed, we show that such a policy is asymptotically optimal in the above sense, if it samples each source with a certain minimum long-run frequency that depends on the source itself and the true subset of anomalous sources.

Our second main result is that this criterion for asymptotic optimality is satisfied, simultaneously for every possible scenario regarding the anomalous sources, when sampling each source at each time instance with a probability that is not smaller than the minimum long-run frequency that corresponds to the current estimate of the subset of anomalous sources. This implies the asymptotic optimality of the *unmodified* policy in [20], as well as of a much simpler policy, according to which the sources are sampled at each time instance conditionally independent of one another given the current estimate of the subset of anomalous sources. Indeed, the implementation of the latter policy, unlike that in [20], involves minimal computational and storage requirements under any setup of the problem. Moreover, we present simulation results that suggest that this computational simplicity does not come at the price of performance deterioration (relative to the policy in [20]).

Finally, to illustrate the gains of asymptotic optimality, we consider the straightforward policy in which the sources are sampled in tandem. We compute its asymptotic relative efficiency and we show that it is asymptotically optimal only in a very special setup of the problem. Moreover, our simulation results suggest that, apart from this special setup, its actual performance loss relative to the above asymptotically optimal policies, when the target error probabilities are not (very) small, is (much) larger than the one implied by its asymptotic relative efficiency.

The remainder of the paper is organized as follows: in Section II we formulate the proposed problem. In Section III we present a family of policies that satisfy the error constraints, as well as an auxiliary consistency property. In Section IV we introduce the family

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

of probabilistic sampling rules, and in Section V we present our asymptotic optimality theory. In Section VI we discuss alternative sampling approaches in the literature, and in Section VII we present the results of our simulation studies. In Section VIII we conclude and discuss potential generalizations, as well as directions for further research. The proofs of all main results are presented in Appendices B, C, D, whereas in Appendix A we state and prove two supporting lemmas.

We end this section with some notations we use throughout the paper. We use $:=$ to indicate the definition of a new quantity and $\equiv$ to indicate a duplication of notation. We set $\mathbb{N} := \{1, 2 \ldots, \}$ and $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$, we denote by A^c the complement, by $|A|$ the size and by 2^A the powerset of a set A, by $\lfloor a \rfloor$ the floor and by $\lceil a \rceil$ the ceiling of a positive number a, and by $\mathbf{1}$ the indicator of an event. We write $x \sim y$ when $\lim(x/y) = 1$, $x \gtrsim y$ when $\lim\inf(x/y) \geq 1$, and $x \lesssim y$ when $\lim\sup(x/y) \leq 1$, where the limit is taken in some sense that will be specified. Moreover, *iid* stands for independent and identically distributed, and we say that a sequence of positive numbers (a_n) is *summable* if $\sum_{n=1}^{\infty} a_n < \infty$ and *exponentially decaying* if there are real numbers $c, d > 0$ such that $a_n \leq c \exp\{-d\, n\}$ for every $n \in \mathbb{N}$. A property that we use in many proofs is that if (a_n) is exponentially decaying, so is the sequence $(\sum_{m \geq \zeta n} a_m)$, for any $\zeta \in (0, 1]$.

II. Problem formulation

Let $(\mathbb{S}, \mathcal{S})$ be an arbitrary measurable space and let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space that hosts M independent sequences of iid $\mathbb{S}$-valued random elements, $\{X_i(n) : n \in \mathbb{N}\}, i \in [M]$, which are generated by M distinct data sources, as well as an independent sequence of iid random vectors, $\{Z(n) : n = 0, 1, \ldots\}$, to be used for randomization purposes. Specifically, each $Z(n) := (Z_0(n), Z_1(n), \ldots, Z_M(n))$ is a vector of independent random variables, uniform in $(0, 1)$, and each $X_i(n)$ has density f_i, with respect to some σ-finite measure ν_i, that is equal to either f_{1i} or f_{0i}. For every $i \in [M]$ we say that source i is "anomalous" if $f_i = f_{1i}$ and we assume that the Kullback-Leibler divergences of f_{1i} and f_{0i} are positive and finite, i.e.,

$$\begin{aligned} I_i &:= \int_{\mathbb{S}} \log(f_{1i}/f_{0i})\, f_{1i}\, d\nu_i \in (0, \infty), \\ J_i &:= \int_{\mathbb{S}} \log(f_{0i}/f_{1i})\, f_{0i}\, d\nu_i \in (0, \infty). \end{aligned} \tag{1}$$

We assume that it is known *a priori* that there are at least ℓ and at most u anomalous sources, where ℓ and u are given, user-specified integers such that $0 \leq \ell \leq u \leq M$, with the understanding that if $\ell = u$, then $0 < \ell < M$. Thus, the family of all possible subsets of anomalous sources is

$$\mathcal{P}_{\ell,u} := \{A \subseteq [M] : \ell \leq |A| \leq u\}.$$

In what follows, we denote by P_A the underlying probability measure and by E_A the corresponding expectation when the subset of anomalous sources is $A \in \mathcal{P}_{\ell,u}$, and we simply write P and E whenever the identity of the subset of anomalous sources is not relevant.

The problem we consider in this work is the identification of the anomalous sources, if any, on the basis of sequentially acquired observations from all sources, when however it is not possible to observe all of them at every sampling instance. Specifically, we have to specify a random time T, at which sampling is terminated, and two random sequences, $R := \{R(n), n \geq 1\}$ and $\Delta := \{\Delta(n), n \in \mathbb{N}\}$, so that $R(n) \subseteq [M]$ represents the subset of sources that are sampled at time n when $n \leq T$, and $\Delta(n) \equiv \Delta_n \in \mathcal{P}_{\ell,u}$ represents the subset of sources that are identified as anomalous when $T = n$.

The decisions whether to stop or not at each time instance, which sources to sample next in the latter case, and which ones to identify as anomalous in the former, must be based on the already available information. Thus, we say that R is a *sampling rule* if, for every $n \in \mathbb{N}$, $R(n)$ is $\mathcal{F}_{n-1}^R$–measurable, where

$$\begin{aligned} \mathcal{F}_n^R &:= \sigma\left(\mathcal{F}_{n-1}^R,\, Z(n),\, \{X_i(n) : i \in R(n)\}\right) \\ \mathcal{F}_0^R &:= \sigma(Z(0)). \end{aligned} \tag{2}$$

Moreover, we say that the triplet (R, T, Δ) is a *policy* if R is a sampling rule, $\{T = n\} \in \mathcal{F}_n^R$ and Δ_n is $\mathcal{F}_n^R$–measurable for every $n \in \mathbb{N}$, in which case we refer to T as *stopping rule* and to Δ as *decision rule*. For any sampling rule R, we denote by $R_i(n)$ the indicator of whether source i is sampled at time n, i.e., $R_i(n) := \mathbf{1}\{i \in R(n)\}$, and by $N_i^R(n)$ (resp. $\pi_i^R(n)$) the number (resp. proportion) of times source i is sampled in the first n time instances, i.e.,

$$N_i^R(n) := \sum_{m=1}^{n} R_i(m), \qquad \pi_i^R(n) := N_i^R(n)/n.$$

We say that a policy (R, T, Δ) belongs to class $\mathcal{C}(\alpha, \beta, \ell, u, K)$ if its probabilities of at least one *false alarm* and at least one *missed detection* upon stopping do not exceed α and β respectively, i.e.,

$$\begin{aligned} \mathsf{P}_A\left(T < \infty,\, \Delta_T \setminus A \neq \emptyset\right) &\leq \alpha \quad \forall\, A \in \mathcal{P}_{\ell,u}, \\ \mathsf{P}_A\left(T < \infty,\, A \setminus \Delta_T \neq \emptyset\right) &\leq \beta \quad \forall\, A \in \mathcal{P}_{\ell,u}, \end{aligned} \tag{3}$$

where α, β are user-specified numbers in $(0, 1)$, and the ratio of its expected total number of observations over its expected time until stopping does not exceed K, i.e.,

$$\sum_{i=1}^{M} \mathsf{E}\left[N_i^R(T)\right] \leq K\, \mathsf{E}[T], \tag{4}$$

where K is a user-specified, real number in $(0, M]$. Note that, in view of the identity

$$\sum_{i=1}^{M} \mathsf{E}\left[N_i^R(T)\right] = \mathsf{E}\left[\sum_{n=1}^{T} \mathsf{E}\left[|R(n)| \mid \mathcal{F}_{n-1}^R\right]\right],$$

constraint (4) is clearly satisfied when

$$\sup_{n \le T} \mathsf{E}\left[|R(n)| \mid \mathcal{F}_{n-1}^R\right] \le K. \tag{5}$$

This is the case, for example, when at most $\lfloor K \rfloor$ sources are sampled at each time instance up to stopping, i.e., when $|R(n)| \le \lfloor K \rfloor$ for every $n \le T$.

Our main goal in this work is, for any given ℓ, u, K, to obtain policies that attain the smallest possible expected time until stopping,

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) := \inf_{(R,T,\Delta) \in \mathcal{C}(\alpha,\beta,\ell,u,K)} \mathsf{E}_A[T], \tag{6}$$

simultaneously under every $A \in \mathcal{P}_{\ell,u}$, to a first-order asymptotic approximation as α and β go to 0. Specifically, when $\ell = u$, α and β are allowed to go to 0 at arbitrary rates, but when $\ell < u$, we assume that

$$\exists\, r \in (0, \infty): \ |\log \alpha| \sim r\,|\log \beta|. \tag{7}$$

III. A FAMILY OF POLICIES

In this section we introduce the statistics that we use in this work, a family of policies that satisfy the error constraint (3), as well as an auxiliary consistency property.

A. Log-likelihood ratio statistics

Let $A, C \in \mathcal{P}_{\ell,u}$ and $n \in \mathbb{N}$. We denote by $\Lambda_{A,C}^R(n)$ the log-likelihood ratio of P_A versus P_C based on the first n time instances when the sampling rule is R, i.e.,

$$\Lambda_{A,C}^R(n) := \log \frac{d\mathsf{P}_A}{d\mathsf{P}_C}\left(\mathcal{F}_n^R\right), \tag{8}$$

and we observe that it admits the following recursion:

$$\begin{aligned} \Lambda_{A,C}^R(n) = \Lambda_{A,C}^R&(n-1) \\ &+ \sum_{i \in A \setminus C} g_i(X_i(n))\, R_i(n) \\ &- \sum_{j \in C \setminus A} g_j(X_j(n))\, R_j(n), \end{aligned} \tag{9}$$

where $\Lambda_{A,C}^R(0) := 0$ and

$$g_i := \log\left(f_{1i}/f_{0i}\right),\ i \in [M]. \tag{10}$$

Indeed, this recursion is obtained by (2) and the facts that $R(n)$ is $\mathcal{F}_{n-1}^R$–measurable, $X_i(n)$ is independent of $\mathcal{F}_{n-1}^R$ and its density under P_A is f_{1i} if $i \in A$ and f_{0i} if $i \notin A$, and $Z(n)$ is independent of both $\mathcal{F}_{n-1}^R$ and $\{X_i(n) : i \in [M]\}$ and has the same density under P_A and P_C.

For any $i, j \in [M]$ we write

$$\begin{aligned} \Lambda_{A,C}^R(n) &\equiv \Lambda_{ij}^R(n) \quad \text{when} \quad A = \{i\}, C = \{j\}, \\ \Lambda_{A,C}^R(n) &\equiv \Lambda_{i}^R(n) \quad \text{when} \quad A = \{i\}, C = \emptyset, \end{aligned}$$

and we observe that the recursion in (9) implies that

$$\Lambda_i^R(n) = \sum_{m=1}^{n} g_i(X_i(m))\, R_i(m), \tag{11}$$

$$\Lambda_{A,C}^R(n) = \sum_{i \in A \setminus C} \Lambda_i^R(n) - \sum_{j \in C \setminus A} \Lambda_j^R(n). \tag{12}$$

In particular, for any $i, j \in [M]$ we have

$$\Lambda_{ij}^R(n) = \Lambda_i^R(n) - \Lambda_j^R(n).$$

In what follows, we refer to $\Lambda_i^R(n)$ as the *local log-likelihood ratio* (LLR) of source i at time n. We introduce the order statistics of the LLRs at time n,

$$\Lambda_{(1)}^R(n) \ge \ldots \ge \Lambda_{(M)}^R(n),$$

and we denote by $w_i^R(n), i \in [M]$ the corresponding indices, i.e.,

$$\Lambda_{(i)}^R(n) := \Lambda_{w_i^R(n)}^R(n), \quad i \in [M].$$

Moreover, we denote by $p^R(n)$ the number of positive LLRs at time n, i.e.,

$$p^R(n) := \sum_{i=1}^{M} \mathbf{1}\{\Lambda_i^R(n) > 0\},$$

and we also set

$$\Lambda_{(0)}^R(n) := +\infty, \qquad \Lambda_{(M+1)}^R(n) := -\infty.$$

B. Stopping and decision rules

We next introduce, for any sampling rule R, a stopping rule, T^R, and a decision rule, Δ^R, such that the policy (R, T^R, Δ^R) satisfies the error constraint (3).

The forms of T^R and Δ^R depend on whether the number of anomalous sources is known in advance or not, i.e., on whether $\ell = u$ or $\ell < u$. Specifically, when $\ell = u$, we stop as soon as the ℓ^{th} largest LLR exceeds the next one by $c > 0$, i.e.,

$$T^R := \inf\left\{n \in \mathbb{N}:\ \Lambda_{(\ell)}^R(n) - \Lambda_{(\ell+1)}^R(n) \ge c\right\}, \tag{13}$$

and we identify as anomalous the sources with the ℓ largest LLRs, i.e.,

$$\Delta_n^R := \left\{w_1^R(n), \ldots, w_\ell^R(n)\right\}, \quad n \in \mathbb{N}. \tag{14}$$

When the number of anomalous sources is completely unknown ($\ell = 0$ and $u = M$), we stop as soon as the value of every LLR is outside $(-a, b)$ for some $a, b > 0$, i.e.,

$$T^R := \inf\left\{n \in \mathbb{N} : \Lambda_i^R(n) \notin (-a, b),\, \forall i \in [M]\right\} \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

and we identify as anomalous the sources with positive LLRs, i.e.,

$$\Delta_n^R := \left\{i \in [M] : \Lambda_i^R(n) > 0\right\}, \quad n \in \mathbb{N}. \tag{16}$$

When $\ell < u$, in general, we combine the stopping rules of the two previous cases and we set

$$\begin{aligned} T^R := \inf \Big\{ & n \in \mathbb{N} : \\ & \Lambda^R_{(\ell+1)}(n) \leq -a \;\;\&\;\; \Lambda^R_{(\ell)}(n) - \Lambda^R_{(\ell+1)}(n) \geq c, \\ & \qquad \text{or} \\ & \ell \leq p^R(n) \leq u \;\;\&\;\; \Lambda_i^R(n) \notin (-a, b) \quad \forall i \in [M], \\ & \qquad \text{or} \\ & \Lambda^R_{(u)}(n) \geq b \;\;\&\;\; \Lambda^R_{(u)}(n) - \Lambda^R_{(u+1)}(n) \geq d \Big\}, \end{aligned} \tag{17}$$

where $a, b, c, d > 0$, and we use the following decision rule:

$$\Delta_n^R := \left\{w_i^R(n) : \; i = 1, \ldots, (p^R(n) \vee \ell) \wedge u\right\} \tag{18}$$

for all $n \in \mathbb{N}$. That is, we identify as anomalous the sources with positive LLRs as long as their number is between ℓ and u. If this number is larger than u (resp. smaller than ℓ), then we declare as anomalous the sources with the u (resp. ℓ) largest LLRs.

Proposition 3.1: Let R be an arbitrary sampling rule.

- When $\ell = u$, (R, T^R, Δ^R) satisfies the error constraint (3) if
$$c = |\log(\alpha \wedge \beta)| + \log(\ell(M - \ell)) \tag{19}$$
- When $\ell < u$, (R, T^R, Δ^R) satisfies the error constraint (3) if
$$\begin{aligned} a &= |\log \beta| + \log M, \\ b &= |\log \alpha| + \log M, \\ c &= |\log \alpha| + \log((M - \ell)M), \\ d &= |\log \beta| + \log(uM). \end{aligned} \tag{20}$$

Proof: When all sources are sampled at all times, this is shown in [6, Theorems 3.1, 3.2]. Essentially the same proof applies when $K < M$, for any sampling rule. ■

In view of Proposition 3.1, in what follows we assume that the thresholds in T^R are selected according to (19) when $\ell = u$ and according to (20) when $\ell < u$. While this is a rather conservative choice, it is sufficient for obtaining asymptotically optimal policies of the form (R, T^R, Δ^R). For this reason, in what follows we say that a sampling rule, R, that satisfies the sampling constraint (4) with $T = T^R$, is *asymptotically optimal under* P_A, for some $A \in \mathcal{P}_{\ell,u}$, if

$$\mathsf{E}_A\left[T^R\right] \sim \mathcal{J}_A(\alpha, \beta, \ell, u, K)$$

as α and β go to 0 at arbitrary rates when $\ell = u$ and so that (7) holds when $\ell < u$. We simply say that R is *asymptotically optimal* if it is asymptotically optimal under P_A for every $A \in \mathcal{P}_{\ell,u}$. We next introduce a weaker property, which will be useful for establishing asymptotic optimality.

C. Exponential consistency

For any sampling rule R and any subset $A \in \mathcal{P}_{\ell,u}$ we denote by σ_A^R the random time starting from which the sources in A are the ones estimated as anomalous by Δ^R, i.e.,

$$\sigma_A^R := \inf\left\{n \in \mathbb{N} : \Delta_m^R = A \quad \text{for all} \;\; m \geq n\right\}, \tag{21}$$

and we say that R is *exponentially consistent under* P_A if $\mathsf{P}_A(\sigma_A^R > n)$ is an exponentially decaying sequence. We simply say that R is *exponentially consistent* if it is exponentially consistent under P_A for every $A \in \mathcal{P}_{\ell,u}$. The following theorem states sufficient conditions for exponential consistency under P_A.

Theorem 3.1: Let $A \in \mathcal{P}_{\ell,u}$ and let R be an arbitrary sampling rule.

- When $\ell < u$, R is exponentially consistent under P_A if there exists a $\rho > 0$ such that $\mathsf{P}_A\left(\pi_i^R(n) < \rho\right)$ is exponentially decaying for every $i \in A$ if $|A| > \ell$ and for every $i \notin A$ if $|A| < u$.
- When $\ell = u$, R is exponentially consistent under P_A if there exists a $\rho > 0$ such that $\mathsf{P}_A\left(\pi_i^R(n) < \rho\right)$ is exponentially decaying either for every $i \in A$ or for every $i \notin A$.

Proof: Appendix B. ■

Remark: Theorem 3.1 reveals that when $|A| = \ell > 0$ or $|A| = u < M$, it is possible to have exponentially consistency under P_A without sampling at all certain sources. Specifically, when $|A| = \ell < u$ (resp. $|A| = u > \ell$) it is not necessary to sample any source in A (resp. A^c). On the other hand, when $|A| = \ell = u$, it suffices to sample either all sources in A or all of them in A^c.

IV. Probabilistic sampling rules

In this section we introduce a family of sampling rules and we show how to design them in order to satisfy the sampling constraint (5) and be exponentially consistent. Thus, we say that a sampling rule R is *probabilistic* if there exists a function

$$q^R : 2^{[M]} \times \mathcal{P}_{\ell,u} \to [0, 1]$$

such that, for every $n \in \mathbb{N}$, $D \in \mathcal{P}_{\ell,u}$, $B \subseteq [M]$,

$$q^R\left(B; D\right) := \mathsf{P}\left(R(n+1) = B \,|\, \mathcal{F}_n^R, \Delta_n^R = D\right), \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

i.e., $q^R(B;D)$ is the probability that B is the subset of sampled sources when D is the currently estimated subset of anomalous sources. If R is a probabilistic sampling rule, then for every $i \in [M]$ and $D \in \mathcal{P}_{\ell,u}$ we set

$$\begin{aligned} c_i^R(D) &:= \mathsf{P}\left(R_i(n+1) = 1 \mid \mathcal{F}_n^R, \Delta_n^R = D\right) \\ &= \sum_{B \subseteq [M]:\, i \in B} q^R(B;D), \end{aligned} \tag{23}$$

i.e., $c_i^R(D)$ is the probability that source i is sampled when D is the currently estimated subset of anomalous sources.

We refer to a probabilistic sampling rule R as *Bernoulli* if, for every $D \in \mathcal{P}_{\ell,u}$ and $B \subseteq [M]$,

$$q^R(B;D) = \prod_{i \in B} c_i^R(D) \prod_{j \notin B} \left(1 - c_j^R(D)\right), \tag{24}$$

i.e., if the sources are sampled at each time instance conditionally independent of one another given the currently estimated subset of anomalous sources. Indeed, such a sampling rule admits a representation of the form

$$R_i(n+1) = \mathbf{1}\left\{Z_i(n) \le c_i^R\left(\Delta_n^R\right)\right\}, \quad n \in \mathbb{N} \tag{25}$$

for all $i \in [M]$, where $Z_1(n), \dots, Z_M(n)$ are iid and uniform in $(0,1)$, thus, its implementation at each time instance requires the realization of M Bernoulli random variables.

The following proposition provides a sufficient condition for a probabilistic sampling rule to satisfy the sampling constraint (5), and consequently (4), for any $\{\mathcal{F}_n^R\}$-stopping time T.

Proposition 4.1: If R is a probabilistic sampling rule such that

$$\sum_{i=1}^M c_i^R(D) \le K \quad \text{for every } D \in \mathcal{P}_{\ell,u}, \tag{26}$$

then (5) holds for any $\{\mathcal{F}_n^R\}$-stopping time T.

Proof: For any $n \in \mathbb{N}$ and any probabilistic sampling rule R,

$$\begin{aligned} \mathsf{E}\left[|R(n)| \mid \mathcal{F}_{n-1}^R\right] &= \sum_{i=1}^M \mathsf{P}\left(R_i(n) = 1 \mid \mathcal{F}_{n-1}^R\right) \\ &= \sum_{i=1}^M c_i^R(\Delta_n^R). \end{aligned} \tag{27}$$

As a result, if (26) is satisfied, then (5) holds for any $\{\mathcal{F}_n^R\}$-stopping time, T.

■

Finally, we establish sufficient conditions for the exponentially consistency of a probabilistic sampling rule.

Theorem 4.1: Let R be a probabilistic sampling rule.

- When $\ell < u$, R is exponentially consistent if, for every $D \in \mathcal{P}_{\ell,u}$, $c_i^R(D)$ is positive for every $i \in D$ if $|D| > \ell$ and every $i \notin D$ if $|D| < u$.
- When $\ell = u$, R is exponentially consistent if, for every $D \in \mathcal{P}_{\ell,u}$, $c_i^R(D)$ is positive either for every $i \in D$ or for every $i \notin D$.

Proof: The proof consists in showing that the sufficient conditions of Theorem 3.1 are satisfied for every $A \in \mathcal{P}_{\ell,u}$, and is presented in Appendix B.

■

Remark: Theorem 4.1 implies that when $\ell < u$, a probabilistic sampling rule is exponentially consistent if, whenever the number of estimated anomalous sources is larger than ℓ (resp. smaller than u), it samples with positive probability any source that is currently estimated as anomalous (resp. non-anomalous). When $\ell = u$, on the other hand, it suffices to sample with positive probability at any time instance either every source that is currently estimated as anomalous or every source that is currently estimated as non-anomalous.

Remark: The proof of Theorem 4.1 relies on two Lemmas, B.1 and B.2, which we state in Appendix B, and is one of the main contributions of this work. We note that the proof would simplify considerably if we strengthened the assumption of the theorem and assumed that every source were sampled with positive probability at every time instance, i.e., if we assumed that $c_i^R(D) > 0$ for every $i \in [M]$ and $D \in \mathcal{P}_{\ell,u}$. Indeed, in this case Lemma B.2 would be redundant and the proof would essentially rely on ideas from [18]. However, such an assumption would unnecessarily and considerably limit the scope of the asymptotic optimality theory we develop in the next section.

V. Asymptotic Optimality

In this section we present the asymptotic optimality theory of this work and discuss some of its implications. For this, we first need to introduce some additional notation.

A. Notation

For $A \subseteq [M]$ with $A \neq \emptyset$ we set

$$\begin{aligned} I_A^* := \min_{i \in A} I_i, \quad I_A := \frac{|A|}{\sum_{i \in A}(1/I_i)} \\ \hat{K}_A := |A| \, \frac{I_A^*}{I_A}, \end{aligned} \tag{28}$$

for $A \subseteq [M]$ with $A \neq [M]$ we set

$$J_A^* := \min_{i \notin A} J_i, \quad J_A := \frac{|A^c|}{\sum_{i \notin A}(1/J_i)} \\ \check{K}_A := |A^c| \, \frac{J_A^*}{J_A}, \tag{29}$$

and for $A \subseteq [M]$ with $0 < |A| < M$ we set

$$\theta_A := \; I_A^*/J_A^*. \tag{30}$$

That is,

- I_A^* is the minimum and I_A the harmonic mean of the Kullback-Leibler divergences in $\{I_i : i \in A\}$,
- J_A^* is the minimum J_A the harmonic mean of the Kullback-Leibler divergences in $\{J_i : i \notin A\}$,
- $\hat{K}_A$ (resp. $\check{K}_A$) is a positive real number smaller or equal to the size of A (resp. A^c), with the equality attained when $I_i = I$ (resp. $J_i = J$) for every i in A (resp. A^c).
- θ_A is a positive real number that is equal to 1 when $I_i = J_i$ for every $i \in [M]$.

In Theorem 5.1, for each $A \in \mathcal{P}_{\ell,u}$ we also introduce two quantities,

$$x_A(r, \ell, u, K) \quad \text{and} \quad y_A(r, \ell, u, K), \tag{31}$$

which play an important role in the formulation of all results in this section. Although their values depend on r, ℓ, u, K, in order to lighten the notation we simply write x_A and y_A unless we want to emphasize their dependence on one or more of these parameters.

B. A universal asymptotic lower bound

Theorem 5.1: Let $A \in \mathcal{P}_{\ell,u}$.

(i) Suppose that $\ell = u$. Then, as $\alpha, \beta \to 0$,

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log(\alpha \wedge \beta)|}{x_A \, I_A^* + y_A \, J_A^*}, \tag{32}$$

where if $\hat{K}_A \leq \theta_A \check{K}_A$,

$$x_A := (K/\hat{K}_A) \wedge 1 \\ y_A := \big((K - \hat{K}_A)^+/\check{K}_A\big) \wedge 1, \tag{33}$$

and if $\hat{K}_A > \theta_A \check{K}_A$,

$$x_A := \big((K - \check{K}_A)^+/\hat{K}_A\big) \wedge 1 \\ y_A := (K/\check{K}_A) \wedge 1. \tag{34}$$

(ii) Suppose that $\ell < u$ and let $\alpha, \beta \to 0$ so that (7) holds.

- If $\ell < |A| < u$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \alpha|}{x_A \, I_A^*} \sim \frac{|\log \beta|}{y_A \, J_A^*} \\ x_A := \frac{K}{\hat{K}_A + (\theta_A/r)\check{K}_A} \wedge (r/\theta_A) \wedge 1 \\ y_A := \frac{K}{\check{K}_A + (r/\theta_A)\hat{K}_A} \wedge (\theta_A/r) \wedge 1. \tag{35}$$

- If $|A| = \ell$, we distinguish two cases.

If $\ell = 0$ or $r \leq 1$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \beta|}{x_A \, I_A^* + y_A \, J_A^*}, \\ \text{where} \quad x_A := 0 \\ y_A := (K/\check{K}_A) \wedge 1. \tag{36}$$

If $\ell > 0$ and $r > 1$, we set

$$z_A := \theta_A/(r - 1)$$

and we distinguish two further cases:

If $z_A \geq 1$ *or* $K \leq \hat{K}_A + z_A \, \check{K}_A$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \beta|}{y_A \, J_A^*} \\ \sim \frac{|\log \alpha|}{x_A \, I_A^* + y_A \, J_A^*}, \tag{37}$$

where

$$x_A := \frac{K}{\hat{K}_A + z_A \, \check{K}_A} \wedge (1/z_A) \wedge 1 \\ y_A := \frac{K}{\check{K}_A + (1/z_A) \, \hat{K}_A} \wedge z_A \wedge 1. \tag{38}$$

If $z_A < 1$ *and* $K > \hat{K}_A + z_A \, \check{K}_A$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \alpha|}{x_A \, I_A^* + y_A \, J_A^*}, \tag{39}$$

where

$$x_A := 1 \\ y_A := \big((K - \hat{K}_A)/\check{K}_A\big) \wedge 1. \tag{40}$$

- If $|A| = u$, then we distinguish again two cases.

If $u = M$ or $r \geq 1$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \alpha|}{x_A \, I_A^* + y_A \, J_A^*} \\ \text{where} \quad y_A := 0 \\ x_A := (K/\hat{K}_A) \wedge 1. \tag{41}$$

If $u < M$ and $r < 1$, we set

$$w_A := (1/\theta_A)/(1/r - 1)$$

and we distinguish two further cases:

If $w_A \geq 1$ or $K \leq \check{K}_A + w_A \hat{K}_A$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \alpha|}{x_A \, I_A^*} \\ \sim \frac{|\log \beta|}{x_A \, I_A^* + y_A \, J_A^*}, \tag{42}$$

where

$$\begin{aligned} y_A &:= \frac{K}{\check{K}_A + w_A \hat{K}_A} \wedge (1/w_A) \wedge 1 \\ x_A &:= \frac{K}{\hat{K}_A + (1/w_A)\check{K}_A} \wedge w_A \wedge 1. \end{aligned} \tag{43}$$

If $w_A < 1$ *and* $K > \check{K}_A + w_A \hat{K}_A$, then

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \gtrsim \frac{|\log \beta|}{x_A\, I_A^* + y_A\, J_A^*}, \tag{44}$$

where

$$\begin{aligned} y_A &:= 1 \\ x_A &:= \big((K - \check{K}_A)/\hat{K}_A\big) \wedge 1. \end{aligned} \tag{45}$$

Proof: The proof is presented in Appendix C. It follows similar steps as the one in the full sampling case in [6, Theorem 5.1], with the difference that it uses a version of Doob's optional sampling theorem in the place of Wald's identity and requires the solution of a max-min optimization problem, in each of the cases we distinguish, to determine the denominator in the lower bound.

■

Remark: By the definition of x_A and y_A in Theorem 5.1 we can see that

- they both take values in $[0, 1]$ and at least one of them is positive,
- they are both increasing as functions of K,
- at least one of them is equal to 1 when $K = M$,
- if $x_A = 0$ (resp. $y_A = 0$), then $|A| = \ell$ (resp. $|A| = u$).

In the next section we obtain an interpretation of the values of x_A and y_A.

C. A criterion for asymptotic optimality

Based on the universal asymptotic lower bound of Theorem 5.1, we next establish the asymptotic optimality under P_A of a sampling rule R that satisfies the sampling constraint and samples each source $i \in [M]$, when the true subset of anomalous sources is A, with a long-run frequency that is not smaller than

$$c_i^*(A) := \begin{cases} x_A\, I_A^*/I_i & \text{if } i \in A, \\ y_A\, J_A^*/J_i & \text{if } i \notin A. \end{cases} \tag{46}$$

Theorem 5.2: Let $A \in \mathcal{P}_{\ell,u}$ and let R be a sampling rule that satisfies (4) with $T = T^R$. If for every $i \in [M]$ such that $c_i^*(A) > 0$ the sequence $\mathsf{P}_A\left(\pi_i^R(n) < \rho\right)$ is summable for every $\rho \in (0, c_i^*(A))$, then R is asymptotically optimal under P_A.

Proof: The proof consists in establishing asymptotic upper bounds for $\mathsf{E}_A[T^R]$, when the thresholds are selected according to (19)-(20), that match the universal asymptotic lower bounds of Theorem 5.1. The proof is presented Appendix D.

■

Remark: For any $A \in \mathcal{P}_{\ell,u}$, by the properties of x_A and y_A and the definition of $\{c_i^*(A) : i \in [M]\}$ in (46) it follows that:

- $c_i^*(A) \in [0, 1]$ for every $i \in [M]$,
- $x_A = 0 \Leftrightarrow c_i^*(A) = 0$ for every $i \in A$,
- $y_A = 0 \Leftrightarrow c_i^*(A) = 0$ for every $i \notin A$,
- $x_A > 0 \Leftrightarrow c_i^*(A) > 0$ for every $i \in A$,
- $y_A > 0 \Leftrightarrow c_i^*(A) > 0$ for every $i \notin A$.

Therefore, Theorem 5.2 implies that when x_A (resp. y_A) is equal to 0, it is not necessary to sample any source in A (resp. A^c) in order to achieve asymptotic optimality under P_A.

Remark: By (46) and the definitions of I_A and J_A in (28) and (29) we can see that

$$\begin{aligned} x_A &= \frac{I_A/I_A^*}{|A|} \sum_{i \in A} c_i^*(A) \quad \text{when} \quad A \neq \emptyset, \\ y_A &= \frac{J_A/J_A^*}{|A^c|} \sum_{i \notin A} c_i^*(A) \quad \text{when} \quad A \neq [M]. \end{aligned} \tag{47}$$

Therefore, x_A (resp. y_A) is equal to the average of the minimum limiting sampling frequencies of the sources in A (resp. A^c) required for asymptotic optimality under P_A, multiplied by I_A/I_A^* (resp. J_A/J_A^*).

Remark: By the definitions of $\hat{K}_A$ and $\check{K}_A$ in (28) and (29), we can see that (47) implies

$$\sum_{i=1}^{M} c_i^*(A) = x_A \hat{K}_A + y_A \check{K}_A. \tag{48}$$

Moreover, by a direct inspection of the values of x_A and y_A we have

$$x_A \hat{K}_A + y_A \check{K}_A \leq K, \tag{49}$$

and consequently:

$$\sum_{i=1}^{M} c_i^*(A) \leq K. \tag{50}$$

D. Asymptotically optimal probabilistic sampling rules

Using the criterion Theorem 5.2, we next obtain a sufficient condition for the asymptotic optimality, simultaneously under every possible scenario, of an arbitrary probabilistic sampling rule, R, in terms of the quantities $\{c_i^R(A) : i \in [M], A \in \mathcal{P}_{\ell,u}\}$, defined in (23).

Theorem 5.3: If R is a probabilistic sampling rule such that, for every $A \in \mathcal{P}_{\ell,u}$,

$$c_i^R(A) \geq c_i^*(A), \quad \forall\, i \in [M], \tag{51}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

then it is exponentially consistent. If also R satisfies (26), then it is asymptotically optimal.

Proof: Exponential consistency is established by showing that the conditions of Theorem 4.1 are satisfied, and asymptotic optimality by showing that the conditions of Theorem 5.2 are satisfied. The proof is presented in Appendix D. ■

In the next corollary we show that both conditions of Theorem 5.3 are satisfied when the equality holds in (51) for every $A \in \mathcal{P}_{\ell,u}$.

Corollary 5.1: If R is a probabilistic sampling rule such that, for every $A \in \mathcal{P}_{\ell,u}$,

$$c_i^R(A) = c_i^*(A), \quad \forall\, i \in [M], \tag{52}$$

then R is exponentially consistent and asymptotically optimal.

Proof: By Theorem 5.3 it clearly suffices to show that (26) holds, which follows directly by (50). ■

While (52) suffices for the asymptotic optimality of a probabilistic sampling rule under P_A, it is not always necessary. In the following proposition we characterize the setups for which the necessity holds.

Proposition 5.1: Let $A \in \mathcal{P}_{\ell,u}$.

(52) holds for any probabilistic sampling rule R that satisfies (26) and (51)

$\Leftrightarrow x_A \hat{K}_A + y_A \check{K}_A = K.$

Proof: For any probabilistic sampling rule R that satisfies (26) and (51) we have

$$K \geq \sum_{i=1}^{M} c_i^R(A) \geq \sum_{i=1}^{M} c_i^*(A) = x_A \hat{K}_A + y_A \check{K}_A, \tag{53}$$

where the equality follows by (48).

($\Leftarrow$) If $K = x_A \hat{K}_A + y_A \check{K}_A$, then by (53) we obtain

$$\sum_{i=1}^{M} c_i^R(A) = \sum_{i=1}^{M} c_i^*(A).$$

In view of (51), this proves that $c_i^R(A) = c_i^*(A)$ for every $i \in [M]$.

($\Rightarrow$) We argue by contradiction and assume that $x_A \hat{K}_A + y_A \check{K}_A = K$ does not hold. By (48) and (49) it then follows that

$$\sum_{i=1}^{M} c_i^*(A) < K.$$

Corollary 5.1 then implies that there is a probabilistic sampling rule that satisfies (51) *with strict inequality for at least one* $i \in [M]$, and also (26). Thus, we have reached a contradiction. ■

E. The optimal performance under full sampling

Corollary 5.1 implies that the asymptotic lower bound in Theorem 5.1 is always achieved, and as a result it is a first-order asymptotic approximation to the optimal expected time until stopping. By this approximation it follows that, for any $A \in \mathcal{P}_{\ell,u}$,

$$\begin{aligned} &\mathcal{J}_A(\alpha,\beta,\ell,u,K) \sim \mathcal{J}_A(\alpha,\beta,\ell,u,M) \\ &\quad \Leftrightarrow \; K \geq Q_A, \end{aligned} \tag{54}$$

where $Q_A \equiv Q_A(r,\ell,u)$ is defined as follows:

$$Q_A := x_A(r,\ell,u,M)\,\hat{K}_A + y_A(r,\ell,u,M)\,\check{K}_A. \tag{55}$$

Moreover, by an inspection of the values of x_A and y_A in Theorem 5.2 it follows that

$$\begin{aligned} Q_A = \min \Big\{ &K \in (0,M] : \\ &x_A(r,\ell,u,K) = x_A(r,\ell,u,M) \\ &\text{and } \; y_A(r,\ell,u,K) = y_A(r,\ell,u,M) \Big\}. \end{aligned} \tag{56}$$

The equivalence in (54) implies that if $Q_A < M$ and $K \in [Q_A, M)$, then the optimal expected time until stopping under P_A *under full sampling* can be achieved, to a first-order asymptotic approximation as $\alpha, \beta \to 0$, *without* sampling all sources at all times.

Corollary 5.2: If $Q_A < M$ and $K \in [Q_A, M)$ for some $A \in \mathcal{P}_{\ell,u}$, then there is a probabilistic sampling rule R such that, for any $\alpha, \beta \in (0,1)$,

$$(R, T^R, \Delta^R) \in \mathcal{C}(\alpha,\beta,\ell,u,K),$$

when the thresholds are selected according to (19)-(20), $c_i^R(A) < 1$ for some $i \in [M]$, and

$$\mathsf{E}_A[T^R] \sim \mathcal{J}_A(\alpha,\beta,\ell,u,M).$$

Proof: This follows by Corollary 5.1 and (54). ■

The next proposition, in conjunction with Proposition 5.1, shows that Q_A can also be used to characterize the setups for which the equality must hold in (51). In particular, it shows that this is always the case when $K \leq 1$.

Proposition 5.2: For every $A \in \mathcal{P}_{\ell,u}$, $Q_A \geq 1$ and

$$x_A \hat{K}_A + y_A \check{K}_A = K \;\Leftrightarrow\; K \leq Q_A.$$

Proof: By the definition of $\hat{K}_A, \check{K}_A$ in (28), (29) it follows that

$$\begin{aligned} \hat{K}_A &= \sum_{i \in A} (I_A^*/I_i) \geq 1 \quad \text{for } A \neq \emptyset \\ \check{K}_A &= \sum_{i \notin A} (J_A^*/J_i) \geq 1 \quad \text{for } A \neq [M]. \end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

Since also at least one of $x_A(r, \ell, u, M)$ and $y_A(r, \ell, u, M)$ is equal to 1, by the definition of Q_A in (55) we conclude that $Q_A \geq 1$. It remains to prove the equivalence. ($\Rightarrow$) It suffices to observe that since x_A, y_A are both increasing in K, by the definition of Q_A in (55) we have $x_A \hat{K}_A + y_A \check{K}_A \leq Q_A$. ($\Leftarrow$) Follows by direct verification. ∎

F. The Chernoff sampling rule

When K is an integer, we will refer to a probabilistic sampling rule as *Chernoff* if it satisfies the conditions of Theorem 5.3 and samples *exactly K sources per time instance*. Indeed, such a sampling rule is implied from [18], [20], [21] when these works are applied to the sequential anomaly detection problem (with a fixed number of sampled sources per time instance). In fact, if the class $\mathcal{C}(\alpha, \beta, \ell, u, K)$ is restricted to policies that sample *exactly* K sources per time instance, the asymptotic optimality of this rule under P_A is implied by the general results in [20], *as long as x_A and y_A are both positive*. However, this is not always the case. Indeed, in the simplest formulation of the sequential anomaly detection problem, where it is known a priori that there is only one anomaly ($l = u = 1$), only one source can be sampled per time instance ($K = 1$), and the sources are homogeneous, i.e., $f_{0i} = f_0$, $f_{1i} = f_1$ for every $i \in [M]$, then one of x_A and y_A is 0 for every $A \in \mathcal{P}_{\ell,u}$. In this setup, the asymptotic optimality of a Chernoff rule was shown in [21, Appendix A] if also f_0, f_1 are both Bernoulli pmf's. Our results in this section remove all these restrictions and establish the asymptotic optimality of the Chernoff rule for any values of ℓ, u, r, K, and without artificially modifying it at a subsequence of sampling instances, as in [25].

From a practical point of view, in order to implement a Chernoff rule one needs to determine a function q^R that satisfies simultaneously the conditions of Theorem 5.3 and also

$$q^R(B; D) = 0 \quad \text{for all } D \in \mathcal{P}_{\ell,u} \quad \text{and} \quad B \subseteq [M] \text{ with } |B| \neq K. \tag{57}$$

This can be a computationally demanding task unless the problem has a special structure or $K = 1$ and should be compared with the implementation of the asymptotically optimal *Bernoulli* sampling rule, which does not require essentially any computation under any setup of the problem.

G. The homogeneous setup

We now specialize the previous results to the case that $I_i = I$ and $J_i = J$ for every $i \in [M]$, where the quantities in (28), (29), (30), (51) simplify as follows:

$$\begin{aligned} & I_A = I_A^* \equiv I, \; J_A = J_A^* \equiv J, \\ & \hat{K}_A = |A|, \; \check{K}_A = |A^c|, \; \theta_A = I/J \equiv \theta, \\ & c_i^*(A) = \begin{cases} x_A, \text{ if } i \in A, \\ y_A, \text{ if } i \notin A. \end{cases} \end{aligned} \tag{58}$$

• Suppose first that the number of anomalous sources is known in advance, i.e., $\ell = u$. Then, x_A and y_A do not depend on A, we denote them simply by x and y respectively, and present their values in Table I.

	$(M-\ell)I \geq J\ell$	$(M-\ell)I \leq J\ell$
x	$\min\{K/\ell, 1\}$	$(K-M+\ell)^+/\ell$
y	$(K-\ell)^+/(M-\ell)$	$\min\{K/(M-\ell), 1\}$

TABLE I: $x \equiv x_A$ and $y \equiv y_A$ when $\ell = u$, $I_i = I$, $J_i = J$ for every $i \in [M]$.

From Table I and (55) it follows that $Q_A = M$ for every $A \in \mathcal{P}_{\ell,u}$. Then, from Propositions 5.1 and 5.2 it follows that if R is a probabilistic sampling rule that satisfies the conditions of Theorem 5.3, then it samples at each time instance each source that is currently estimated as anomalous (resp. non-anomalous) with probability x (resp. y), i.e.,

$$c_i^R(A) = \begin{cases} x, & \text{if } i \in A \\ y, & \text{if } i \notin A, \end{cases} \quad \forall\, A \in \mathcal{P}_{\ell,u}. \tag{59}$$

Moreover, we observe that the first-order asymptotic approximation to the optimal performance is independent of the true subset of anomalous sources. Specifically, for every $A \in \mathcal{P}_{\ell,u}$,

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \sim \frac{|\log(\alpha \wedge \beta)|}{x\, I + y\, J}. \tag{60}$$

• When the number of anomalous sources is not known a priori, i.e., $\ell < u$, we focus on the special case that $r = 1$ and $\theta = 1$, or equivalently, $I = J$. Then, the values of x_A and y_A, presented in Table II, do not depend on I, and the optimal asymptotic performance under P_A takes the following form:

$$\begin{aligned} & I\, \frac{\mathcal{J}_A(\alpha, \beta, \ell, u, K)}{|\log \alpha|} \\ & \sim \begin{cases} \max\{(M-\ell)/K, 1\}, & |A| = \ell, \\ M/K, & \ell < |A| < u, \\ \max\{u/K, 1\}, & |A| = u. \end{cases} \end{aligned} \tag{61}$$

From (54) and (61) we further obtain

$$Q_A = \begin{cases} M-\ell, & \text{if} \quad |A| = \ell, \\ M, & \text{if} \quad \ell < |A| < u, \\ u, & \text{if} \quad |A| = u. \end{cases} \tag{62}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

Note that, in this setup, $Q_A = M$ if and only if one of the following holds:

$$\ell < |A| < u, \quad |A| = \ell = 0, \quad |A| = u = M.$$

Moreover, from Theorem 5.2 it follows that, in each of these three cases, asymptotic optimality is achieved by any sampling rule, not necessarily probabilistic, that satisfies the sampling constraint and samples all sources with the same long-run frequency. This is the content of the following corollary.

Corollary 5.3: Let R be a sampling rule such that

- the sampling constraint (4) holds with $T = T^R$,
- $\mathsf{P}\left(|\pi_i^R(n) - K/M| > \epsilon\right)$ is summable for every $\epsilon > 0$ and $i \in [M]$.

If $\ell < u$, $r = 1$, and $I_i = J_j$ for every $i, j \in [M]$, then R is asymptotically optimal under P_A for every $A \in \mathcal{P}_{\ell,u}$ with $\ell < |A| < u$. If also $\ell = 0$ and $u = M$, then R is asymptotically optimal.

Proof: This follows directly by Theorem 5.2, in view of Table II. ■

The conditions of Corollary 5.3 are satisfied when R is a probabilistic sampling rule with $c_i^R(A) = K/M$ for every $i \in [M]$ and $A \in \mathcal{P}_{\ell,u}$, e.g., when R is a Bernoulli sampling rule that samples each source at each time instance with probability K/M, independently of the other sources. Moreover, in the setup of Corollary 5.3 it is quite convenient to find and implement a Chernoff rule (Subsection V-F). Indeed, when K is an integer, the conditions of Corollary 5.3 are satisfied when we take a *simple random sample* of K sources at each time instance, i.e., when

$$\begin{aligned} q^R(B; D) = 1/\binom{M}{K} \quad &\text{for all } \; D \in \mathcal{P}_{\ell,u} \\ &\text{and } B \subseteq [M] \;\text{ with }\; |B| = K. \end{aligned} \tag{63}$$

Finally, in Subsection VI-A we will introduce a non-probabilistic sampling rule that satisfies the conditions of Corollary 5.3.

H. A heterogeneous example

We end this section by considering a setup where M is an even number, $\ell < M/2 < u$, $r = 1$, and

$$I_i = J_i = \begin{cases} I, & 1 \le i \le M/2, \\ I/\phi & M/2 < i \le M, \end{cases} \tag{64}$$

for some $\phi \in (0, 1]$. Moreover, we focus on the case that the subset of anomalous sources is of the form $A = \{1, \ldots, |A|\}$. Then, the optimal asymptotic performance under P_A takes the following form:

- when $|A| = \ell$,
$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \sim \frac{|\log \alpha|}{I/\phi} \max\left\{\frac{(\phi+1)M/2 - \ell}{K}, 1\right\}, \tag{65}$$
- when $\ell < |A| < u$,
$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \sim \frac{|\log \alpha|}{I} \max\left\{\frac{(\phi+1)M/2}{K}, 1\right\}, \tag{66}$$
- when $|A| = u$,
$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \sim \frac{|\log \alpha|}{I} \max\left\{\frac{(1-\phi)(M/2) + \phi u}{K}, 1\right\}. \tag{67}$$

From these expressions and (54) we also obtain

$$Q_A = \begin{cases} (\phi+1)M/2 - \ell, & |A| = \ell, \\ (\phi+1)M/2, & \ell < |A| < u, \\ (1-\phi)(M/2) + \phi u, & |A| = u, \end{cases} \tag{68}$$

and we note that Q_A is always strictly smaller than M in this setup as long as $\phi < 1$.

VI. Non-probabilstic sampling rules

In this section, we discuss certain non-probabilistic sampling rules.

A. Sampling in tandem

Suppose that K is an integer and consider the straightforward sampling approach according to which the sources are sampled in tandem, K of them at a time. Specifically, sources 1 to K are sampled at time $n = 1$, and if $2K \le M$, then sources $K+1$ to $2K$ are sampled at time $n = 2$, whereas if $2K > M$, then sources $K+1$ to M and 1 to $2K - M$ are sampled at time $n = 2$, etc. In this way, each source is sampled exactly K times in an interval of the form $((m-1)M, mM]$, where $m \in \mathbb{N}$, to which we refer as a *cycle*. This sampling rule satisfies the conditions of Corollary 5.3, which means that in the special case that $\ell < u$, $r = 1$, and $I_i = J_j$ for every $i, j \in [M]$, it achieves asymptotic optimality under P_A when $\ell < |A| < u$, and for every $A \subseteq [M]$ when $\ell = 0$ and $u = M$.

In general, by the formula for the optimal asymptotic performance under full sampling, which is obtained by the lower bound of Theorem 5.1 with $K = M$, we can see that if sampling is terminated at a time that is a multiple of M, the expected *number of cycles* until stopping is, to a first-order asymptotic approximation, equal to $\mathcal{J}(\alpha, \beta, \ell, u, M)/K$. Since each cycle is of length M, the expected *time* until stopping is, again to a first-order asymptotic approximation, equal to

	$\lvert A\rvert=\ell=0$	$\lvert A\rvert=u=M$	$\ell<\lvert A\rvert<u$	$0<\ell=\lvert A\rvert<u$	$\ell<\lvert A\rvert=u<M$
x_A	0	K/M	K/M	0	$\min\{K/u,1\}$
y_A	K/M	0	K/M	$\min\{K/(M-\ell),1\}$	0

TABLE II: x_A and y_A when $\ell<u$, $r=1$, $I_i=J_j$ for every $i,j\in[M]$.

$(M/K)\mathcal{J}(\alpha,\beta,\ell,u,M)$. Thus, the *asymptotic relative efficiency* of this sampling approach can be defined as follows:

$$\mathsf{ARE} := \frac{M}{K}\lim_{\alpha,\beta\to 0}\frac{\mathcal{J}_A(\alpha,\beta,\ell,u,M)}{\mathcal{J}_A(\alpha,\beta,\ell,u,K)}, \tag{69}$$

where the limit is taken so that (7) holds when $\ell<u$.

Consider in particular the homogeneous setup of Subsection V-G, where (58) holds. When $\ell=u$, by (59) and (60) we have:

- if $\theta\geq \ell/(M-\ell)$,
$$\mathsf{ARE} = \frac{\theta}{1+\theta}\,\frac{M}{\max\{\ell,K\}}+\frac{1}{1+\theta}\frac{(1-\ell/K)^+}{1-\ell/M}, \tag{70}$$
- if $\theta< \ell/(M-\ell)$,
$$\begin{aligned}\mathsf{ARE} =&\frac{1}{1+\theta}\,\frac{M}{\max\{M-\ell,K\}}\\ &+\frac{\theta}{1+\theta}\frac{(1-(M-\ell)/K)^+}{\ell/M}.\end{aligned} \tag{71}$$

When $\ell<u$ and $r=\theta=1$, by (61) we obtain:

$$\mathsf{ARE} = \begin{cases} M/\max\{M-\ell,K\}, & \lvert A\rvert=\ell,\\ 1, & \ell<\lvert A\rvert<u,\\ M/\max\{u,K\} & \lvert A\rvert=u.\end{cases} \tag{72}$$

On the other hand, in the heterogeneous setup of Subsection V-H, by (65), (66), (67) we obtain:

$$\mathsf{ARE} = \begin{cases} M/\max\{(\phi+1)M/2-\ell,\ K\}, & \lvert A\rvert=\ell,\\ M/\max\{(\phi+1)M/2,\ K\}, & \ell<\lvert A\rvert<u,\\ M/\max\{(1-\phi)M/2+\phi u,\ K\}, & \lvert A\rvert=u.\end{cases}$$

B. Equalizing empirical and limiting sampling frequencies

We next consider a sampling approach, which has been applied to a general controlled sensing problem [28], as well as to a bandit problem [29], and we show that not only it may not achieve asymptotic optimality in the sequential anomaly detection problem, but it may even lead to a detection procedure that *fails to terminate with positive probability*.

To be more specific, we consider the homogeneous setup of Subsection V-G with $K=1$. In this setup, a probabilistic sampling rule that satisfies the conditions of Theorem 5.3 samples a source in D (resp. D^c) with probability x_D (resp. y_D), whenever $D\in\mathcal{P}_{\ell,u}$ is the subset of sources currently identified as anomalous. The sampling rule R that we consider in this Subsection is not probabilistic, as it uses not only the currently estimated subset of anomalous sources, but also of the current empirical sampling frequencies. Specifically, if D is the subset of sources currently estimated as anomalous, for every source in D (resp. D^c) it computes the distance between its current empirical sampling frequency and x_D (resp. y_D), and it samples next a source for which this distance is the maximum. That is, for every $n\in\mathbb{N}$ and $D\in\mathcal{P}_{\ell,u}$, $R(n+1)$ is on the event $\{\Delta_n^R=D\}$ a subset of

$$\begin{aligned}\text{argmax}\Big\{&\lvert\pi_i^R(n)-x_D\rvert,\\ &\lvert\pi_j^R(n)-y_D\rvert\,:\, i\in D,\, j\notin D\Big\}.\end{aligned} \tag{73}$$

Without any loss of generality, we also assume that each source has positive probability to be sampled at the first time instance, i.e.,

$$\mathsf{P}_A(i\in R(1))>0\quad\forall\, i\in[M],\ A\in\mathcal{P}_{\ell,u}. \tag{74}$$

Proposition 6.1: Consider the homogeneous setup of Subsection V-G with $K=1$ and let R be sampling rule that satisfies (73)-(74). Suppose further that there is only one anomalous source, i.e., that the subset of anomalous source, A, is a singleton, and also that $x_A+y_A<1$. Then, there is an event of positive probability under P_A on which

(i) the same source is sampled at every time instance,
(ii) and if also $\ell=0$ and $u=M$, T^R fails to terminate for any selection of its thresholds.

Proof: If (i) holds, there is an event of positive probability under P_A on which all LLRs but one are always equal to 0. Thus, (ii) follows directly from (i) and the fact that when $\ell=0$ and $u=M$, the stopping rule (15) requires that all LLRs be non-zero upon stopping. Therefore, it remains to prove (i).

Without loss of generality, we set $A=\{1\}$. Moreover, we recall the definition of g_1 in (10) and define the event

$$\Gamma := \left\{\sum_{m=1}^{n} g_1(X_1(m))>0\quad\forall\, n\in\mathbb{N}\right\}. \tag{75}$$

Since $\{g_1(X_1(n))\,:\, n\in\mathbb{N}\}$ is an iid sequence with expectation $I>0$ under P_A (see, e.g., [30, Proposition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

7.2.1]), Γ has positive probability under P_A. This is also the case for $\{R_1(1) = 1\}$, due to (74). Since these two events are independent, their intersection also has positive probability under P_A. Therefore, it suffices to show that only source 1 is sampled on $\Gamma \cap \{R_1(1) = 1\}$. Indeed, on this event sampling starts from source 1 and, as a result, the vector of empirical frequencies, $(\pi_1^R(n), \ldots, \pi_M^R(n))$ at time $n = 1$ is $(1, 0, \ldots, 0)$. Moreover, the estimated subset of anomalous sources at $n = 1$ is $A = \{1\}$, independently of whether $\ell < u$ or $\ell = u$. Since $x_A + y_A < 1$, or equivalently,

$$\begin{aligned} |\pi_1(1) - x_A| &= |1 - x_A| \\ &= 1 - x_A \\ &> y_A = |y_A - 0| = |\pi_i(1) - y_A| \end{aligned} \tag{76}$$

for every $i \neq 1$, by (73) it follows that source 1 is sampled again at time $n = 2$ and, as a result, the vector of empirical frequencies remains $(1, 0, \ldots, 0)$ at time $n = 2$. Applying the same reasoning as before, we conclude that source 1 is sampled again at time $n = 3$. The same argument can be repeated indefinitely, and this proves (i).

■

Remark: When $\ell = 0$, $u = M$, each f_{0i} is exponential with rate 1, and each f_{1i} is exponential with rate $\lambda > 0$, the conditions of the previous proposition are satisfied as long as $M > 2$. Indeed, in this case we have

$$\theta = I/J = \frac{-\log(\lambda) + \lambda - 1}{\log(\lambda) + 1/\lambda - 1},$$
$$x_A = \frac{1}{1 + (M-1)\theta}, \quad y_A = \theta x_A,$$

and, as a result, $x_A + y_A < 1 \Leftrightarrow M > 2$.

C. Sampling based on the ordering of the LLRs

A different, non-probabilistic sampling approach, which goes back to [18, Remark 5], suggests sampling at each time instance the sources with the smallest, in absolute value, LLRs among those estimated as anomalous/non-anomalous. Such a sampling rule was proposed in [12], in the homogeneous setup of Subsection V-G, under the assumption that the number of anomalous sources is known a priori, i.e., $\ell = u$. An extension of this rule in the heterogeneous setup was studied in [13], under the assumption that $\ell = u$ and $K = 1$. A similar sampling rule, that also has a randomization feature, was proposed in [16] when $\ell = u$, as well as in the case of no prior information, i.e., when $\ell = 0, u = M$. For each of them, the criterion of Theorem 5.2 can be applied to establish their asymptotic optimality. Its verification, however, is a quite difficult task that we plan to consider in the future.

VII. Simulation study

In this section we present the results of a simulation study in which we compare two probabilistic sampling rules, Bernoulli (Section IV) and Chernoff (Subsection V-F), between them and against sampling in tandem (Subsection VI-A). Throughout this section, for every $i \in [M]$ we have $f_{0i} = \mathcal{N}(0, 1)$ and $f_{1i} = \mathcal{N}(\mu_i, 1)$, i.e., all observations from source i are Gaussian with variance 1 and mean equal to μ_i if the source is anomalous and 0 otherwise, and as a result $I_i = J_i = (\mu_i)^2/2$. We consider a homogeneous setup where

$$\mu_i = \mu, \quad \forall\, i \in [M], \tag{77}$$

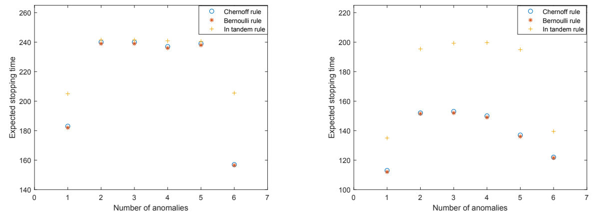
as well as a heterogeneous setup where

$$\mu_i = \begin{cases} \mu, & 1 \leq i \leq M/2 \\ 2\mu, & M/2 < i \leq M, \end{cases} \tag{78}$$

in which case (64) holds with $I = \mu^2/2$ and $\phi = 0.25$. In both setups we set $\alpha = \beta$, $M = 10$, $K = 5$, $\ell = 1$, $u = 6$, $\mu = 0.5$, we assume that the subset of anomalous sources is of the form $A = \{1, \ldots, |A|\}$, and consider different values for its size. The two probabilistic sampling rules are designed so that (52) holds for every $A \in \mathcal{P}_{\ell,u}$. As a result, by Corollary 5.1 it follows that, in both setups, they are asymptotically optimal under P_A for every $A \in \mathcal{P}_{\ell,u}$. On the other hand, by Corollary 5.3 implies that sampling in tandem is asymptotically optimal under P_A only in the homogeneous setup (78) and when $l < |A| < u$, since $0 < l < u < M$.

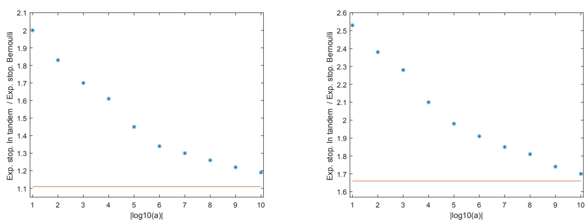
In Figure 1 we plot the expected value of the stopping time that is induced by each of the three sampling rules against the true number of anomalous sources. Specifically, in Figure 1a we consider the homogeneous setup (77) and in Figure 1b the heterogeneous setup (78). In all cases, the thresholds are chosen, via Monte Carlo simulation, so that the familywise error probability of each kind is (approximately) equal to $\alpha = \beta = 10^{-3}$. The Monte Carlo standard error that corresponds to each estimated expected value is approximately 10^{-2}. From Figure 1 we can see that the performance implied by the two probabilistic sampling rules is always essentially the same. Sampling in tandem performs significantly worse in all cases apart from the homogeneous setup with $\ell < |A| < u$, where all three sampling rules lead to essentially the same performance. We note also that, in both setups and for all three sampling rules, the expected time until stopping is much smaller when the number of anomalous sources is equal to either ℓ or u than when it is between ℓ and u.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory



(a) Homogeneous case (b) Heterogeneous case

Fig. 1: Expected value of the stopping time that corresponds to each of the three sampling rules versus the number of anomalous sources ($\alpha = \beta = 10^{-3}$).



(a) $|A| = 1$ (b) $|A| = 6$

Fig. 2: In both graphs the x-axis represents $|\log_{10}(\alpha)|$ and the y-axis the ratio of the expected value of the stopping time implied by sampling in tandem over that implied by the asymptotically optimal Bernoulli rule in the homogeneous setup (77). The horizontal line refers to the value of (72).

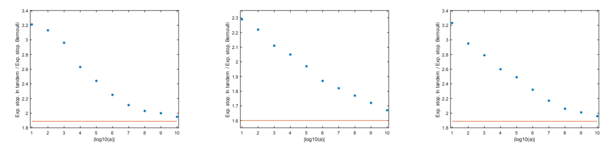
In Figures 2 and 3 we plot the ratio of the expected value of the stopping time induced by sampling in tandem over that induced by the Bernoulli sampling rule against $|\log_{10}(\alpha)|$ as α ranges from 10^{-1} to 10^{-10}. (We do not present the corresponding results for the Chernoff rule, as they are almost identical). Specifically, in Figure 2 we consider the homogeneous setup (77) when the number of anomalous sources is 1 and 6, whereas in Figure 3 we consider the heterogeneous setup (78) when the number of anomalous sources is 1, 3, and 6. For each value of α, the thresholds are selected according to (20) and each expectation is computed using 10^4 simulation runs. The standard error for each estimated expectation is approximately 1, whereas the standard error for each ratio is approximately 10^{-2} in the homogeneous setup and $3 \cdot 10^{-2}$ in the heterogeneous setup. Moreover, in each case we plot the limiting value of this ratio, which is the limit defined in (69). In the homogeneous case this is given by (72) and is equal to

$$\begin{cases} 10/\max\{10-1,\ 5\} \approx 1.1, & |A| = 1, \\ 10/\max\{6,\ 5\} \approx 1.6, & |A| = 6. \end{cases} \tag{79}$$

In the heterogeneous case it is given by (VI-A) and is equal to

$$\begin{cases} 10/\max\{1.25 \cdot 5 - 1,\ 5\} \approx 1.9, & |A| = 1, \\ 10/\max\{1.25 \cdot 5,\ 5\} \approx 1.6, & 1 < |A| < 6, \\ 10/\max\{0.75 \cdot 5 + 0.25 \cdot 6,\ 5\} \approx 1.9, & |A| = 6. \end{cases}$$

From Figures 2 and 3 we can see that, in all cases, the efficiency loss due to sampling in tandem is (much) larger than the one suggested by the corresponding asymptotic relative efficiency when α is not (very) small.



(a) $|A| = 1$ (b) $|A| = 3$ (c) $|A| = 6$

Fig. 3: In all graphs the x-axis represents $|\log_{10}(\alpha)|$ and the y-axis the ratio of the expected value of the stopping time implied by sampling in tandem over that implied by the asymptotically optimal Bernoulli rule in the heterogeneous setup (78). The horizontal line refers to the value of (VI-A).

VIII. Conclusions and extensions

In this paper we propose a novel formulation of the sequential anomaly detection problem with sampling constraints, in which arbitrary, user-specified bounds are assumed on the number of anomalous sources, the probabilities of at least one false alarm and at least one missed detection are controlled below distinct tolerance levels, and the number of sampled sources per time instance is not necessarily fixed. We obtain a general criterion for achieving the optimal expected time until stopping, to a first-order asymptotic approximation as the error probabilities go to 0, as long as the log-likelihood ratio statistic of each observation has a finite *first* moment. We show that asymptotic optimality is achieved, simultaneously under every possible subset of anomalous sources, for any version of the proposed problem, using the *unmodified* sampling rule in [20], to which we refer as *Chernoff*, but also using a novel sampling rule whose implementation requires minimal computations, to which we refer as *Bernoulli*. Despite their very different computational requirements, these two rules are found in simulation studies to lead to essentially the same performance.

In various works in the relevant literature, such as [12], [13], [16], it has been shown, in simulation studies under various setups, that non-probabilistic sampling rules as the ones discussed in Subsection VI-C can lead

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

to significantly better performance in practice compared to the Chernoff rule. The study of sampling rules of this nature in the general sequential anomaly detection framework we propose in this work is an interesting problem that we plan to consider in the future. Another interesting problem is that of establishing stronger asymptotic optimality properties, as it was done in [23], [24] for certain special cases of the sequential multi-hypothesis testing problem with controlled sensing.

The results of this paper can be shown, using the techniques in [8], to remain valid for a variety of error metrics beyond the familywise error rates that we consider in this work, such as the false discovery rate and the false non-discovery rate. However, this is not the case for the generalized familywise error rates proposed in [31], [32], for which different policies and a different analysis is required. These error metrics have been considered in [4], [5], [7] in the case of full sampling, whereas certain results in the presence of sampling constraints have been presented in [33].

The results in this work can also be generalized in a natural way when the sampling cost varies per source, as in [15], or when the two hypotheses in each source are not completely specified, as it is done for example in [7] in the case of full sampling. Another potential generalization is the removal of the assumption that the acquired observations are conditionally independent of the past given the current sampling choice, as it is done in [27] in a general controlled sensing setup. Finally, another direction of interest is a setup where the focus is on the dependence structure of the sources rather than their marginal distributions, as for example in [34].

IX. Acknowledgments

The authors would like to thank Prof. Venugopal Veeravalli for stimulating discussions.

Appendix A

In this Appendix we state and prove two auxiliary lemmas that are used in the proofs of various results of this paper. Specifically, we fix $A \in \mathcal{P}_{\ell,u}$, and for any any sampling rule, R, $i \in [M]$, and $n \in \mathbb{N}$, we set

$$\begin{aligned}\widetilde{\Lambda}_i^R(n) &:= \widetilde{\Lambda}_i^R(n-1) \\ &\quad + \Big(g_i(X_i(n)) - \mathsf{E}_A[g_i(X_i(n))]\Big)\, R_i(n), \qquad (80)\\ \widetilde{\Lambda}_i^R(0) &:= 0,\end{aligned}$$

and comparing with (11) we observe that

$$\widetilde{\Lambda}_i^R(n) = \begin{cases} \Lambda_i^R(n) - I_i\, N_i^R(n), & i \in A \\ \Lambda_i^R(n) + J_i\, N_i^R(n), & i \notin A. \end{cases} \qquad (81)$$

Lemma A.1: Let R be an arbitrary sampling rule, $i \in A$, $j \notin A$, $\rho \in (0,1]$, and $\epsilon > 0$. Then, the sequences

$$\begin{aligned}&\mathsf{P}_A\left(\frac{1}{n}\Lambda_i^R(n) < \rho I_i - \epsilon,\ \pi_i^R(n) \geq \rho\right),\\ &\mathsf{P}_A\left(\frac{1}{n}\Lambda_j^R(n) > -\rho J_j + \epsilon,\ \pi_j^R(n) \geq \rho\right),\\ &\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}^R(n) < \rho I_i - \epsilon,\ \pi_i^R(n) \geq \rho\right)\\ &\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}^R(n) < \rho J_j - \epsilon,\ \pi_j^R(n) \geq \rho\right)\end{aligned}$$

are exponentially decaying.

Proof: We prove the result for the third probability, as the proofs for the other ones are similar. To lighten the notation, we suppress the dependence on R and we write $\widetilde{\Lambda}_{ij}(n), \widetilde{\Lambda}_i(n), \pi_i(n), \mathcal{F}_n$ instead of $\widetilde{\Lambda}_{ij}^R(n), \widetilde{\Lambda}_i^R(n), \pi_i^R(n), \mathcal{F}_n^R$. By (81), for every $n \in \mathbb{N}$ we have

$$\begin{aligned}\Lambda_{ij}(n) &= \Lambda_i(n) - \Lambda_j(n)\\ &= \widetilde{\Lambda}_i(n) - \widetilde{\Lambda}_j(n) + n(I_i\pi_i(n) + J_j\pi_j(n)),\end{aligned}$$

which shows that if $\pi_i(n) \geq \rho$, then

$$\Lambda_{ij}(n) \geq \widetilde{\Lambda}_i(n) - \widetilde{\Lambda}_j(n) + n\rho I_i$$

Thus, for every $n \in \mathbb{N}$,

$$\begin{aligned}&\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}(n) < \rho I_i - \epsilon,\ \pi_i(n) \geq \rho\right)\\ &\leq \mathsf{P}_A\left(\widetilde{\Lambda}_i(n) - \widetilde{\Lambda}_j(n) < -n\,\epsilon\right)\\ &\leq \mathsf{P}_A\left(\widetilde{\Lambda}_i(n) < -n\,\epsilon/2\right) + \mathsf{P}_A\left(-\widetilde{\Lambda}_j(n) < -n\,\epsilon/2\right),\end{aligned}$$

and it suffices to show that the two terms in the upper bound are exponentially decaying. We show this only for the first one, as the proof for the second is similar. For this, we fix $\delta \in (0, \epsilon/2)$ and we observe that, for any $t > 0$ and $n \in \mathbb{N}$, by Markov's inequality we have

$$\begin{aligned}&\mathsf{P}_A\left(\widetilde{\Lambda}_i(n) < -n\,\epsilon/2\right)\\ &\leq \exp\{-n\,(\epsilon/2)\,t\}\,\mathsf{E}_A\left[\exp\left\{-t\,\widetilde{\Lambda}_i(n)\right\}\right].\end{aligned} \qquad (82)$$

By (80) and the law of iterated expectation it follows that the expectation in the upper bound can be written as follows:

$$\begin{aligned}&\mathsf{E}_A\Bigg[\exp\left\{-t\,\widetilde{\Lambda}_i(n-1)\right\}\\ &\cdot \mathsf{E}_A\left[\exp\left\{-t\,(g_i(X_i(n)) - I_i)\,R_i(n)\right\} \mid \mathcal{F}_{n-1}\right]\Bigg].\end{aligned} \qquad (83)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

Since $R(n)$ is an $\mathcal{F}_{n-1}$-measurable Bernoulli random variable and $X_i(n)$ is independent of $\mathcal{F}_{n-1}$ and has the same distribution as $X_i(1)$,

$$\begin{aligned}&\mathsf{E}_A\left[\exp\left\{-t\left(g_i(X_i(n))-I_i\right)R_i(n)\right\}\mid\mathcal{F}_{n-1}\right]\\&=\mathsf{E}_A\left[\exp\left\{-t\left(g_i(X_i(1))-I_i\right)\right\}\right]^{R_i(n)}\\&=\mathsf{E}_A\left[\exp\left\{t\left(-g_i(X_i(1))+I_i-\delta+\delta\right)\right\}\right]^{R_i(n)}\\&\leq\exp\left\{R_i(n)\left(\psi_i(t)+t\,\delta\right)\right\},\end{aligned}$$

where ψ_i is the cumulant generating function of $-g_i(X_i(1))+I_i-\delta$, i.e.,

$$\psi_i(t):=\log\mathsf{E}_A\left[\exp\left\{t(-g_i(X_i(1))+I_i-\delta)\right\}\right],\ t>0.$$

Since ψ_i is convex on $(0,1)$, $\psi_i(1)=I_i-\delta<\infty$, and

$$\psi_i'(0+)=\mathsf{E}_A[-g_i(X_i(1))+I_i-\delta]=-\delta<0,$$

there is an $s>0$ such that $\psi_i(s)<0$, and as a result

$$\exp\left\{R_i(n)\left(\psi_i(s)+s\,\delta\right)\right\}\leq\exp\left\{s\,\delta\right\}.$$

Therefore, setting $t=s$ in (82)-(83) we obtain

$$\begin{aligned}&\mathsf{P}_A\big(\widetilde{\Lambda}_i(n)<-n\,\epsilon/2\big)\\&\qquad\leq\exp\{-n(\epsilon/2)s+\delta s\}\cdot\\&\qquad\quad\cdot\mathsf{E}_A\left[\exp\left\{-s\widetilde{\Lambda}_i(n-1)\right\}\right].\end{aligned}\tag{84}$$

Repeating the same argument $n-1$ times we conclude that there exists an $s>0$ such that

$$\mathsf{P}_A\left(\widetilde{\Lambda}_i(n)<-n\epsilon/2\right)\leq\exp\left\{-n(\epsilon/2-\delta)s\right\}.$$

Since $\delta\in(0,\epsilon/2)$, this completes the proof. ∎

Lemma A.2: Let $i\in A$, $j\notin A$, $\zeta>0$, $\rho_i,\rho_j\in[0,1]$, and let R be an arbitrary sampling rule.

(i) If $\rho_i>0$ and $\mathsf{P}_A(\pi_i^R(n)<\rho_i)$ is summable, then so is
$$\mathsf{P}_A\left(\frac{1}{n}\Lambda_i^R(n)<\rho_i I_i-\zeta\right).$$

(ii) If $\rho_j>0$ and $\mathsf{P}_A(\pi_j^R(n)<\rho_j)$ is summable, then so is
$$\mathsf{P}_A\left(\frac{1}{n}\Lambda_j^R(n)>-\rho_j J_j+\zeta\right).$$

(iii) If $\rho_i+\rho_j>0$ and both $\mathsf{P}_A(\pi_j^R(n)<\rho_i)$ and $\mathsf{P}_A(\pi_j^R(n)<\rho_j)$ are summable, then so is
$$\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}^R(n)<\rho_i I_i+\rho_j J_j-\zeta\right).$$

Proof: For (i), by the law of total probability we have

$$\begin{aligned}&\mathsf{P}_A\left(\frac{1}{n}\Lambda_i^R(n)<\rho_i I_i-\zeta\right)\\&\leq\mathsf{P}_A\left(\pi_i^R(n)<\rho_i\right)\\&\quad+\mathsf{P}_A\left(\frac{1}{n}\Lambda_i^R(n)<\rho_i I_i-\zeta,\,\pi_i^R(n)\geq\rho_i\right).\end{aligned}$$

The first term in the upper bound is summable by assumption, whereas the second one is summable, as exponentially decaying, by Lemma A.1.

The proof of (ii) is similar and is omitted. Since (iii) follows directly by (i) and (ii) when ρ_i and ρ_j are both positive, it suffices to consider the case where only one them is positive. Without loss of generality, we assume that $\rho_i>0=\rho_j$. Then, by the law of total probability again we have

$$\begin{aligned}&\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}^R(n)<\rho_i I_i+\rho_j J_j-\zeta\right)\\&\leq\mathsf{P}_A\left(\pi_i^R(n)<\rho_i\right)\\&\quad+\mathsf{P}_A\left(\frac{1}{n}\Lambda_{ij}^R(n)<\rho_i I_i-\zeta,\,\pi_i^R(n)\geq\rho_i\right).\end{aligned}$$

The first term in the upper bound is summable by assumption, whereas the second one is summable, as exponentially decaying, by Lemma A.1. ∎

Appendix B

In this Appendix we prove Theorems 3.1 and 4.1, which provide sufficient conditions for the exponential consistency of a sampling rule. In order to lighten the notation, throughout this Appendix we suppress dependence on R and we write $\pi_i(n)$, $\Lambda_i(n),\Lambda_{ij}(n),\Delta_n$, $c_i(A),\sigma_A$ instead of $\pi_i^R(n)$, $\Lambda_i^R(n)$, $\Lambda_{ij}^R(n)$, Δ_n^R, $c_i^R(A)$, σ_A^R.

Proof of Theorem 3.1: We prove the result first when $\ell=u$. By the definition of the decision rule in (14) it follows that when $\sigma_A>n$, there are $m\geq n$, $i\in A$, $j\notin A$ such that $\Lambda_{ij}(m)\leq 0$. As a result, by the union bound we have

$$\mathsf{P}_A(\sigma_A>n)\leq\sum_{i\in A,\,j\notin A}\sum_{m=n}^{\infty}\mathsf{P}_A\left(\Lambda_{ij}(m)\leq 0\right).$$

Therefore, to show that $\mathsf{P}_A(\sigma_A>n)$ is exponentially decaying, it suffices to show that this is the case for $\mathsf{P}_A\left(\Lambda_{ij}(n)\leq 0\right)$ for every $i\in A$ and $j\notin A$. To show this, we fix such i and j and we note that, by assumption, either $\mathsf{P}_A(\pi_i(n)<\rho)$ or $\mathsf{P}_A(\pi_j(n)<\rho)$ is exponentially decaying for $\rho>0$ small enough. Without loss of generality, suppose that this is the case for the former. By an application of the law of total probability we then obtain

$$\begin{aligned}\mathsf{P}_A\left(\Lambda_{ij}(n)\leq 0\right)\leq\;&\mathsf{P}_A(\Lambda_{ij}(n)\leq 0,\pi_i(n)\geq\rho)\\&+\mathsf{P}_A(\pi_i(n)<\rho).\end{aligned}$$

As mentioned earlier, the second term in the upper bound is, by assumption, exponentially decaying for $\rho>0$ small enough. By Lemma A.1 it follows that this is also the case for the first one.

We next prove the result when $\ell < u$, and consider first the case that $\ell < |A| < u$. By the definition of the decision rule in (18) it follows that when $\sigma_A > n$, there is an $m \geq n$ such that either $\Lambda_i(m) < 0$ for some $i \in A$, or $\Lambda_j(m) > 0$ for some $j \notin A$. As a result, by the union bound we have

$$\begin{aligned} \mathsf{P}_A(\sigma_A > n) \leq & \sum_{i \in A} \sum_{m=n}^{\infty} \mathsf{P}_A(\Lambda_i(m) < 0) \\ & + \sum_{j \notin A} \sum_{m=n}^{\infty} \mathsf{P}_A(\Lambda_j(m) \geq 0). \end{aligned}$$

Therefore, to prove that $\mathsf{P}_A(\sigma_A > n)$ is exponentially decaying, it suffices to show that this is true for $\mathsf{P}_A(\Lambda_i(n) < 0)$ and $\mathsf{P}_A(\Lambda_j(n) \geq 0)$ for every $i \in A$ and $j \notin A$. This can be shown similarly to the case $\ell = u$, in view of the fact that the sequences $\mathsf{P}_A(\pi_i(n) < \rho)$ and $\mathsf{P}_A(\pi_j(n) < \rho)$ are, by assumption, both exponentially decaying for $\rho > 0$ small enough and for every $i \in A$ and $j \notin A$.

The two remaining cases are $\ell < |A| = u$ and $\ell = |A| < u$. We consider only the former, as the proof for the latter is similar, and assume that $\ell < |A| = u$. By the definition of the decision rule in (14) it follows that when $\sigma_A > n$, there is an $m \geq n$ such that either $\Lambda_{ij}(m) \leq 0$ for some $i \in A$ and $j \notin A$, or $\Lambda_i(m) < 0$ for some $i \in A$. As a result, by the union bound we have

$$\begin{aligned} \mathsf{P}_A(\sigma_A > \zeta n) \leq & \sum_{i \in A} \sum_{m=n}^{\infty} \mathsf{P}_A(\Lambda_i(m) < 0) \\ & + \sum_{i \in A, j \notin A} \sum_{m=n}^{\infty} \mathsf{P}_A(\Lambda_{ij}(m) \leq 0). \end{aligned}$$

Since $|A| > \ell$, the sequence $\mathsf{P}_A(\pi_i(n) < \rho)$ is, by assumption, exponentially decaying for every $i \in A$ and $\rho > 0$ small enough. Therefore, similarly to the previous cases we can show that each term in the upper bound is exponentially decaying.

■

In the remainder of this Appendix we prove Theorem 4.1, whose proof relies on two preliminary lemmas. To state those, for every $D \in \mathcal{P}_{\ell,u}$ and $n \in \mathbb{N}$ we denote by τ_n^D the first time instance m at which D has been estimated as the subset of anomalous sources for at least $2\zeta m$ times since $\lceil n/2 \rceil$, i.e.,

$$\begin{aligned} \tau_n^D := \inf \Bigg\{ & m \geq \lceil n/2 \rceil : \\ & \sum_{u=\lceil n/2 \rceil}^{m} \mathbf{1}\{\Delta_{u-1} = D\} \geq 2\zeta m \Bigg\}, \end{aligned} \tag{85}$$

where $\zeta > 0$ is an arbitrary constant, which in the proof of Theorem 4.1 will be selected to be small enough.

The first preliminary lemma intuitively says that if a source $i \in [M]$ has positive probability to be sampled whenever a subset $D \in \mathcal{P}_{\ell,u}$ is estimated as the one containing the anomalous sources, then it is unlikely that in the long run both D will be frequently estimated as the anomalous subset and source i will be infrequently sampled. Note that this lemma does not require any of the conditions of Theorem 4.1.

Lemma B.1: Let $D \in \mathcal{P}_{\ell,u}$ and $i \in [M]$. If $c_i(D) > 0$, then, for any $\rho > 0$ small enough, the sequences

$$\begin{aligned} & \mathsf{P}(\tau_n^D \leq n, \; \pi_i(\tau_n^D) < \rho) \\ & \mathsf{P}(\tau_n^D \leq n, \; \pi_i(n) < \rho) \end{aligned}$$

are exponentially decaying.

Proof: We prove the two claims together by showing that $\mathsf{P}(\tau_n^D \leq n, \; \pi_i(\sigma_n) < \rho)$ is exponentially decaying for all $\rho > 0$ small enough, where σ_n stands for either n or τ_n^D. For any given $n \in \mathbb{N}$ we set

$$\begin{aligned} \widetilde{\pi}_i(n) := & \frac{1}{n} \sum_{m=1}^{n} (R_i(m) - c_i(\Delta_{m-1})) \\ = & \pi_i(n) - \frac{1}{n} \sum_{m=1}^{n} c_i(\Delta_{m-1}), \quad n \in \mathbb{N}. \end{aligned}$$

On the event $\{\tau_n^D \leq n\}$ we have $n/2 \leq \tau_n^D \leq \sigma_n \leq n$ and, as a result,

$$\begin{aligned} \pi_i(\sigma_n) - \widetilde{\pi}_i(\sigma_n) = & \frac{1}{\sigma_n} \sum_{m=1}^{\sigma_n} c_i(\Delta_{m-1}) \\ \geq & \frac{c_i(D)}{n} \sum_{m=\lceil n/2 \rceil}^{\tau_n^D} \mathbf{1}\{\Delta_{m-1} = D\} \\ \geq & \frac{c_i(D)}{n} \, 2\,\zeta\, \tau_n^D \geq \; c_i(D)\,\zeta, \end{aligned}$$

where in the last inequality we have used the definition of τ_n^D. Consequently, for every $\rho > 0$ and $n \in \mathbb{N}$,

$$\begin{aligned} & \{\tau_n^D \leq n, \pi_i(\sigma_n) \leq \rho\} \\ & \qquad \subseteq \{\tau_n^D \leq n, \, \widetilde{\pi}_i(\sigma_n) \leq \rho - c_i(D)\,\zeta\}. \end{aligned}$$

Since $c_i(D) > 0$, there is an $\epsilon > 0$ such that, for all $\rho \in (0, \, c_i(D)\zeta - \epsilon)$,

$$\{\tau_n^D \leq n, \; \pi_i(\sigma_n) \leq \rho\} \subseteq \left\{ \max_{\lceil n/2 \rceil \leq m \leq n} |\widetilde{\pi}_i(m)| \geq \epsilon \right\}.$$

Therefore, it suffices to show that, for all $\epsilon > 0$, the sequence

$$\mathsf{P}\left(\max_{\lceil n/2 \rceil \leq m \leq n} |\widetilde{\pi}_i(m)| \geq \epsilon \right) \tag{86}$$

is exponentially decaying. This follows by the fact that $(R_i(n) - c_i(\Delta_{n-1}))$ is a uniformly bounded martingale difference, in view of (23), and an application of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

the maximal Azuma-Hoeffding submartingale inequality (see, e.g., [35, Remark 2.3]).

■

The second lemma intuitively says that the conditions of Theorem 4.1 imply that, in the long run, it will be unlikely to frequently identify, incorrectly, a subset whose size is either ℓ or u as the one containing the anomalous sources.

Lemma B.2: Let $A, D \in \mathcal{P}_{\ell,u}$ be such that $|D| \in \{l, u\}$ and $D \neq A$. If R is a sampling rule that satisfies the conditions of Theorem 4.1, then the sequence $\mathsf{P}_A(\tau_n^D \leq n)$ is exponentially decaying.

Proof: We consider first the case that $\ell < u$, where we only prove the result when $|D| = u$, as the proof when $|D| = \ell$ is similar. Since $A \in \mathcal{P}_{\ell,u}$, $|D| = u$ and $D \neq A$, there exists a $j \in D \setminus A$. Moreover, since $|D| = u > \ell$, the assumptions of Theorem 4.1 imply that $c_j(D) > 0$. Thus, by Lemma B.1 it follows that

$$\mathsf{P}_A\left(\tau_n^D \leq n,\ \pi_j(\tau_n^D) < \rho\right) \tag{87}$$

is an exponentially decaying sequence for $\rho > 0$ small enough. It remains to show that this is also the case for

$$\mathsf{P}_A\left(\tau_n^D \leq n,\ \ \pi_j(\tau_n^D) \geq \rho\right). \tag{88}$$

Indeed, by the definition of τ_n^D in (85) it follows that on the event $\{\tau_n^D < \infty\}$ we have $\Delta(\tau_n^D - 1) = D$. Since $|D| = u$ and $j \in D$, by the definition of the decision rule in (18) it follows that $\Lambda_j(\tau_n^D - 1) > 0$. Therefore,

$$\begin{aligned}
&\mathsf{P}_A\left(\tau_n^D \leq n,\ \ \pi_j(\tau_n^D) \geq \rho\right) \\
&= \mathsf{P}_A(\tau_n^D \leq n,\ \Lambda_j(\tau_n^D - 1) > 0,\ \pi_j(\tau_n^D) \geq \rho) \\
&\leq \sum_{m=\lceil n/2 \rceil}^{n} \mathsf{P}_A\left(\Lambda_j(m-1) > 0,\ \pi_j(m) \geq \rho\right),
\end{aligned}$$

where the inequality holds because τ_n^D takes values in $[n/2, n]$ on the event $\{\tau_n^D \leq n\}$. Therefore, it remains to show that the sequence

$$\mathsf{P}_A\left(\Lambda_j(n-1) > 0,\ \pi_j(n) \geq \rho\right) \tag{89}$$

is exponentially decaying for $\rho > 0$ small enough. Indeed, for large enough n, $\pi_j(n) \geq \rho$ implies that

$$\pi_j(n-1) \geq \frac{n\pi_j(n) - 1}{n-1} \geq \frac{\rho n - 1}{n-1} \geq \rho - \frac{1}{n-1}.$$

For any given $\rho > 0$, there exists a $\rho' > 0$ so that for all n large enough we have

$$\rho - \frac{1}{n-1} > \rho', \tag{90}$$

and so that the probability in (89) is bounded by

$$\mathsf{P}_A\left(\Lambda_j(n-1) > 0,\ \pi_j(n-1) \geq \rho'\right).$$

For $\rho > 0$ small enough, $\rho' > 0$ is small enough, and by Lemma A.1 it follows that the latter probability, and consequently (89), is exponentially decaying in n.

It remains to prove the lemma when $\ell = u$ and $A \neq D$. In this case there are $i \in A \setminus D$ and $j \in D \setminus A$, and by the assumptions of Theorem 4.1 it follows that either $c_i(D) > 0$ or $c_j(D) > 0$. Without loss of generality, we assume that the latter holds. Then, by Lemma B.1 it follows that (87) is exponentially decaying for all $\rho > 0$ small enough, and it suffices to show that this is also the case for (88). Indeed, by the definition of τ_n^D in (85) it follows that on the event $\{\tau_n^D < \infty\}$ we have $\Delta(\tau_n^D - 1) = D$. Consequently, by the definition of the decision rule in (14) it follows that there is an $i \in A \setminus D$ such that $\Lambda_{ij}(\tau_n^D - 1) < 0$. Therefore, by the union bound we have

$$\begin{aligned}
&\mathsf{P}_A\left(\tau_n^D \leq n,\ \ \pi_j(\tau_n^D) \geq \rho\right) \\
&= \sum_{i \in A \setminus D} \mathsf{P}_A\left(\tau_n^D \leq n,\ \Lambda_{ij}(\tau_n^D - 1) < 0,\ \pi_j(\tau_n^D) \geq \rho\right) \\
&\leq \sum_{i \in A \setminus D} \sum_{m=\lceil n/2 \rceil}^{n} \mathsf{P}_A\left(\Lambda_{ij}(m-1) < 0,\ \pi_j(m) \geq \rho\right),
\end{aligned}$$

where as before the inequality holds because τ_n^D takes values in $[n/2, n]$ on the event $\{\tau_n^D \leq n\}$. Therefore, it remains to show that the sequence

$$\mathsf{P}_A\left(\Lambda_{ij}(n-1) < 0,\ \pi_j(n) \geq \rho\right)$$

is exponentially decaying for $\rho > 0$ small enough. As before, this follows by an application of Lemma A.1.

■

Proof of Theorem 4.1: Fix $A \in \mathcal{P}_{\ell,u}$. By Theorem 3.1 it suffices to show that, for all $\rho > 0$ small enough, $\mathsf{P}_A(\pi_i(n) < \rho)$ is an exponentially decaying sequence

- for every $i \in A$, if $|A| > \ell$, and for every $i \notin A$, if $|A| < u$, when $\ell < u$,
- either for every $i \in A$ or for every $i \notin A$, when $\ell = u$.

In order to do so, we select the positive constant ζ in (85) to be smaller than $1/(4|\mathcal{P}_{\ell,u}|)$. Then, for every $n \in \mathbb{N}$ there is at least one $D \in \mathcal{P}_{\ell,u}$ for which $\{\tau_n^D \leq n\} \neq \emptyset$. As a result, for every $i \in [M]$ and $\rho > 0$, by the union bound we have

$$\mathsf{P}_A(\pi_i(n) < \rho) \leq \sum_{D \in \mathcal{P}_{\ell,u}} \mathsf{P}_A\left(\pi_i(n) < \rho,\ \tau_n^D \leq n\right).$$

Suppose first that $\ell < u$. Then, it suffices to show that, for every $D \in \mathcal{P}_{\ell,u}$ and all $\rho > 0$ small enough,

$$\mathsf{P}_A\left(\pi_i(n) < \rho, \tau_n^D \leq n\right) \tag{91}$$

is exponentially decaying for every $i \in A$ when $|A| > \ell$ and for every $i \notin A$ when $|A| < u$. We only consider the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

former case, as the proof for the latter is similar. Thus, suppose that $|A| > \ell$ and let $i \in A$.

- If $c_i(D) > 0$, by Lemma B.1 it follows that (91) is an exponentially decaying sequence.
- If $c_i(D) = 0$, the assumption of the theorem implies that either $|D| = u$ and $i \notin D$, or $|D| = \ell$ and $i \in D$. In either case, $A \neq D$ and by Lemma B.2 it follows that $\mathsf{P}_A(\tau_n^D \leq n)$, and consequently (91), is an exponentially decaying sequence.

Suppose now that $\ell = u$. Then, it suffices to show that (91) is exponentially decaying for every $D \in \mathcal{P}_{\ell,u}$ and all $\rho > 0$ small enough, either for every $i \in A$ or for every $i \notin A$.

- When $D \neq A$, by Lemma B.2 it follows that $\mathsf{P}_A(\tau_n^D \leq n)$, and consequently (91), is exponentially decaying.
- When $D = A$, then by assumption $c_i(A) > 0$ holds either for every $i \in A$ or for every $i \notin A$. By Lemma B.1 it then follows that (91) is exponentially decaying either for every $i \in A$ or for every $i \notin A$, and this completes the proof.

■

Appendix C

In this Appendix we fix $A \in \mathcal{P}_{\ell,u}$ and prove the universal asymptotic lower bound of Theorem 5.2. The proof relies on two lemmas, for the statement of which we need to introduce the function

$$\phi(\alpha,\beta) := \alpha \log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right),$$

where $\alpha, \beta \in (0,1)$, i.e., the Kullback-Leibler divergence between a Bernoulli distribution with parameter α and one with parameter $1-\beta$. Moreover, we set $\phi(\alpha) \equiv \phi(\alpha,\alpha)$.

The first lemma states a non-asymptotic, information-theoretic inequality that generalizes the one used in Wald's universal lower bound for the problem of testing two simple hypotheses [36, p. 156].

Lemma C.1: Suppose that $\alpha+\beta < 1$ and let (R, T, Δ) be a policy that satisfies the error constraint (3) and $\mathsf{P}_A(T < \infty) = 1$. Then, for any $C \in \mathcal{P}_{\ell,u}$ such that $C \neq A$,

$$\begin{aligned} &\mathsf{E}_A\left[\Lambda^R_{A,C}(T)\right] \\ &\quad \geq \begin{cases} \phi(\alpha,\beta) & \text{if } C\backslash A \neq \emptyset,\ A\backslash C = \emptyset, \\ \phi(\beta,\alpha) & \text{if } C\backslash A = \emptyset,\ A\backslash C \neq \emptyset, \\ \phi(\alpha\wedge\beta) & \text{if } C\backslash A \neq \emptyset,\ A\backslash C \neq \emptyset. \end{cases} \end{aligned} \tag{92}$$

Proof: The proof is identical to that in the full sampling case in [6, Theorem 5.1], and can be obtained by an application of the data processing inequality for Kullback-Leibler divergences (see, e.g., [37, Lemma 3.2.1]). Indeed, the left-hand side is the Kullback-Leibler divergence between P_A and P_C given the available information up to time T, when the sampling rule R is utilized, whereas the right hand side is obtained by considering the Kullback-Leibler divergence between P_A and P_C based on a single event of $\mathcal{F}_T^R$. ■

We next make use of the previous inequality to establish lower bounds on the expected number of samples taken from each source until stopping.

Lemma C.2: Suppose that $\alpha+\beta < 1$ and let (R, T, Δ) be a policy that satisfies the error constraint (3) and $\mathsf{E}_A[T] < \infty$.

(i) If $|A| < u$, then

$$\min_{j\notin A}\left(J_j\, \mathsf{E}_A\left[N_j^R(T)\right]\right) \geq \phi(\alpha,\beta). \tag{93}$$

(ii) If $|A| > \ell$, then

$$\min_{i\in A}\left(I_i\, \mathsf{E}_A\left[N_i^R(T)\right]\right) \geq \phi(\beta,\alpha). \tag{94}$$

(iii) If either $|A| = \ell > 0$ or $|A| = u < M$, then

$$\begin{aligned} &\min_{i\in A}\left(I_i\, \mathsf{E}_A\left[N_i^R(T)\right]\right) \\ &\quad + \min_{j\notin A}\left(J_j\, \mathsf{E}_A\left[N_j^R(T)\right]\right) \geq \phi(\alpha\wedge\beta). \end{aligned} \tag{95}$$

Proof: We recall the sequence $\widetilde{\Lambda}_i^R$, defined in (80), and note that it is a zero-mean, $\{\mathcal{F}_n^R\}$-martingale under P_A. Moreover, by the finiteness of the Kullback-Leibler divergences in (1) we have:

$$\sup_{n\in\mathbb{N}} \mathsf{E}_A\left[|\widetilde{\Lambda}_i^R(n) - \widetilde{\Lambda}_i^R(n-1)| \mid \mathcal{F}_{n-1}^R\right] < \infty.$$

Since also T is an $\{\mathcal{F}_n^R\}$-stopping time such that $\mathsf{E}_A[T] < \infty$, by the Optional Sampling Theorem [38, pg. 251] we obtain:

$$\mathsf{E}_A\left[\widetilde{\Lambda}_i^R(T)\right] = 0 \quad \text{for every} \quad i \in [M]. \tag{96}$$

(i) If $|A| < u$, there is a $j \notin A$ so that the set $C = A \cup \{j\}$ belongs to $\mathcal{P}_{\ell,u}$. By representation (12) and decomposition (81) we also have

$$\Lambda^R_{A,C}(T) = -\Lambda^R_j(T) = -\widetilde{\Lambda}^R_j(T) + J_j\, N_j^R(T),$$

and by (92) and (96) we obtain

$$J_j\, \mathsf{E}_A\left[N_j^R(T)\right] \geq \phi(\alpha,\beta).$$

Since this inequality holds for every $j \notin A$, it proves (93).

(ii) The proof is similar to (i) and is omitted.

(iii) If $|A| = \ell > 0$ or $|A| = u < M$, then there are $i \in A$ and $j \notin A$ so that $C = A \cup \{j\} \setminus \{i\} \in \mathcal{P}_{\ell,u}$. By representation (12) and decomposition (81) we have

$$\begin{aligned} \Lambda^R_{A,C}(T) &= \Lambda^R_i(T) - \Lambda^R_j(T) \\ &= \widetilde{\Lambda}^R_i(T) - \widetilde{\Lambda}^R_j(T) + J_j\, N_j^R(T) + I_i\, N_i^R(T), \end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

and by (92) and (96) we obtain

$$I_i \, \mathsf{E}_A \left[N_i^R(T)\right] + J_j \, \mathsf{E}_A \left[N_j^R(T)\right] \geq \phi(\alpha \wedge \beta).$$

Since this inequality holds for every $i \in A$ and $j \notin A$, it proves (95). ■

For the proof of Theorem 5.1 we introduce the following notation:

$$\mathcal{D}_K := \left\{ (c_1, \ldots, c_M) \in [0,1]^M : \sum_{i=1}^{M} c_i \leq K \right\}, \qquad (97)$$
$$\mathcal{D}'_K := \{(p,q) \in [0,1]^2 : \; p\hat{K}_A + q\check{K}_A \leq K\}.$$

Moreover, we observe that as $\alpha, \beta \to 0$ we have

$$\phi(\alpha, \beta) \sim |\log \beta|, \qquad (98)$$
$$r(\alpha, \beta) \equiv \frac{\phi(\beta, \alpha)}{\phi(\alpha, \beta)} \sim \frac{|\log \alpha|}{|\log \beta|}. \qquad (99)$$

Proof of Theorem 5.1: (i) Let $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$ and $(R, T, \Delta) \in \mathcal{C}(\alpha, \beta, \ell, u, K)$ such that $\mathsf{E}_A[T] < \infty$. By Lemma C.2(iii) it then follows that

$$\mathsf{E}_A[T] \; W_A(T) \geq \phi(\alpha \wedge \beta)$$

where

$$W_A(T) := \min_{i \in A} \left\{ I_i \, \frac{\mathsf{E}_A[N_i^R[T]]}{\mathsf{E}_A[T]} \right\} + \min_{j \notin A} \left\{ J_j \, \frac{\mathsf{E}_A[N_j^R[T]]}{\mathsf{E}_A[T]} \right\}$$

and by constraint (4) we conclude that

$\mathsf{E}_A[T] \; V_A \geq \phi(\alpha \wedge \beta)$,

where $V_A := \max_{(c_1, \ldots, c_M) \in \mathcal{D}_K} \left\{ \min_{i \in A}(c_i I_i) + \min_{j \notin A}(c_j J_j) \right\}.$

Since the lower bound is independent of the policy (R, T, D), we have

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \; V_A \geq \phi(\alpha \wedge \beta).$$

Comparing with (32) and recalling (98), we can see that it suffices to show that $V_A = x_A I_A^* + y_A J_A^*$ with x_A and y_A as in (33)-(34). Indeed, the maximum in V_A is achieved by c_i's of the form

$$c_i I_i = p I_A^*, \quad i \in A,$$
$$c_j J_j = q J_A^*, \quad j \notin A, \qquad (100)$$

for $p, q \in [0,1]$ such that the constraint in $\mathcal{D}_K$ is satisfied, i.e.,

$$K \geq \sum_{i=1}^{M} c_i = p \sum_{i \in A} \frac{I_A^*}{I_i} + q \sum_{j \notin A} \frac{J_A^*}{J_j} = p\,\hat{K}_A + q\,\check{K}_A, \qquad (101)$$

and as a result,

$$V_A = \max_{(p,q) \in \mathcal{D}'_K} \{p I_A^* + q J_A^*\}.$$

This maximum is achieved by $p, q \in [0,1]$ such that $p\hat{K}_A + q\check{K}_A = K \wedge (\hat{K}_A + \check{K}_A)$, in particular by p and q equal to x_A and y_A as in (33)-(34), which completes the proof.

(ii) Suppose first that $\ell < |A| < u$. As before, let $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$ and $(R, T, \Delta) \in \mathcal{C}(\alpha, \beta, \ell, u, K)$ such that $\mathsf{E}_A[T] < \infty$. Then, by Lemma C.2(i) and Lemma C.2(ii) we obtain:

$$\mathsf{E}_A[T] \; W_A(T) \geq \phi(\beta, \alpha),$$

where $W_A(T)$ is now defined as

$$W_A(T) := \min \left\{ \min_{i \in A} \left\{ I_i \, \frac{\mathsf{E}_A[N_i^R[T]]}{\mathsf{E}_A[T]} \right\}, \; r(\alpha, \beta) \, \min_{j \notin A} \left\{ J_j \, \frac{\mathsf{E}_A[N_j^R[T]]}{\mathsf{E}_A[T]} \right\} \right\},$$

and by constraint (4) we conclude that

$$\mathsf{E}_A[T] \; V_A(\alpha, \beta) \geq \phi(\beta, \alpha), \qquad (102)$$

where

$$V_A(\alpha, \beta) := \max_{(c_1, \ldots, c_M) \in \mathcal{D}_K} \min \left\{ \min_{i \in A} (c_i I_i), \; r(\alpha, \beta) \, \min_{j \notin A} (c_j J_j) \right\}. \qquad (103)$$

Since the lower bound does not depend on the policy (R, T, Δ), we further have

$$\mathcal{J}_A(\alpha, \beta, \ell, u, K) \; V_A(\alpha, \beta) \geq \phi(\beta, \alpha).$$

Comparing with (35) and recalling (98), we can see that it suffices to show that

$$V_A(\alpha, \beta) \to x_A \, I_A^* = r \, y_A \, J_A^* \qquad (104)$$

as $\alpha, \beta \to 0$ according to (7), with x_A and y_A as in (35).

The equality in (104) follows directly from the values of x_A and y_A in (37). Moreover, the maximum in $V_A(\alpha, \beta)$ is achieved by $c_1, \ldots, c_M$ of the form (100) that satisfy (101). Therefore:

$$V_A(\alpha, \beta) = \max_{(p,q) \in \mathcal{D}'_K} \min \{p I_A^*, \, r(\alpha, \beta) \, q J_A^*\},$$

and this maximum is achieved for p and q such that the two terms in the minimum are equal. As a result,

$$V_A(\alpha, \beta) = p I_A^* = r(\alpha, \beta) \, q J_A^*,$$

where p and q are equal to x_A and y_A in (35), with r replaced by $r(\alpha, \beta)$. As α and β go to 0 according to (7), we have $r(\alpha, \beta) \to r$ (recall (99)),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

and consequently $V_A(\alpha,\beta) \to x_A\, I_A^*$ with x_A as in (37).

Finally, we consider the case $|A| = \ell$ and omit the proof when $|A| = u$, as it is similar. When either $\ell = 0$ or $r \leq 1$, we have to show (36). Indeed, working as before, using Lemma C.2(i) we obtain

$$\mathcal{J}_A(\alpha,\beta,\ell,u,K)\; V_A \geq \phi(\alpha,\beta),$$
$$\text{where} \quad V_A := \max_{(c_1,\dots,c_M)\in\mathcal{D}_K} \; \min_{j\notin A} \{c_j J_j\}.$$

Comparing with (36), and recalling (98), we can see that it suffices to show that $V_A = J_A^*\, y_A$, with y_A as in (36). Indeed, the maximum in V_A is achieved by $c_1,\dots,c_M$ of the form (100) with $p = 0$ and $q \in [0,1]$ such that (101) is satisfied, i.e.,

$$V_A = J_A^* \max_{q\in[0,1]:\; q\check{K}_A \leq K} q.$$

This shows that $V_A = J_A^*\, y_A$, with y_A as in (36), and completes the proof in this case.

It remains to establish the asymptotic lower bound when $\ell > 0$ and $r > 1$, in which case we have to show (37)-(39). The asymptotic equivalence in (37) can be shown by direct evaluation, therefore it suffices to show only the asymptotic lower bound in this case.

Working as before, using Lemma C.2(i) and Lemma C.2(iii), for any $\alpha,\beta \in (0,1)$ such that $\alpha+\beta < 1$ we obtain

$$\mathcal{J}_A(\alpha,\beta,\ell,u,K)\; V_A(\alpha,\beta) \geq \phi(\beta,\alpha), \tag{105}$$

where

$$V_A(\alpha,\beta) := \max_{(c_1,\dots,c_M)\in\mathcal{D}_K} \min\Big\{ r(\alpha,\beta)\, \min_{j\notin A}(c_j J_j), \; \min_{i\in A}(c_i I_i) + \min_{j\notin A}(c_j J_j) \Big\}. \tag{106}$$

This maximum is achieved by $c_1,\dots,c_M$ of the form (100) that satisfy (101), thus,

$$V_A(\alpha,\beta) = \max_{(p,q)\in\mathcal{D}'_K} \min\{ r(\alpha,\beta)\, qJ_A^*,\; pI_A^* + qJ_A^* \}.$$

If either $\theta_A \geq r(\alpha,\beta) - 1$ or

$$K \leq \hat{K}_A + (\theta_A/(r(\alpha,\beta)-1))\check{K}_A,$$

the maximum in $V_A(\alpha,\beta)$ is achieved when the two terms in the minimum are equal and, as a result,

$$V_A(\alpha,\beta) = p\, I_A^* + q\, J_A^* = r(\alpha,\beta)\; qJ_A^* \tag{107}$$

with p and q equal to x_A and y_A as in (37), again with r replaced by $r(\alpha,\beta)$.

Otherwise, the second term in the minimum is smaller and the first equality in (107) holds with p and q equal to x_A and y_A as in (39), but with r replaced by $r(\alpha,\beta)$.

Therefore, letting α and β go to 0 in (105), according to (7), and recalling (98)-(99), proves the asymptotic lower bounds in both (37) and (39).

■

Appendix D

In this Appendix we fix $A \in \mathcal{P}_{\ell,u}$ and prove Theorems 5.2 and 5.3, which provide sufficient conditions for asymptotic optimality. In both proofs we recall that $x_A \vee y_A > 0$, $x_A > 0$ when $|A| > \ell$, $y_A > 0$ when $|A| < u$, and $c_i^*(A) > 0$ for every i in A (resp. A^c) when $x_A > 0$ (resp. $y_A > 0$).

Proof of Theorem 5.2: We prove the theorem first when $\ell = u$, where α and β go to 0 at arbitrary rates. By the asymptotic lower bound (32) in Theorem 5.1 it follows that, in this case, it suffices to show that

$$\mathsf{E}_A[T^R] \lesssim \frac{|\log(\alpha\wedge\beta)|}{x_A\, I_A^* + y_A\, J_A^*}. \tag{108}$$

To show this, for any $\epsilon > 0$ small enough and any $c > 0$ we set

$$L_\epsilon(c) := \max_{i\in A, j\notin A} \frac{c}{(c_i^*(A)-\epsilon)I_i + (c_j^*(A)-\epsilon)J_j - \epsilon}, \tag{109}$$

and observe that

$$\mathsf{E}_A[T^R] \leq L_\epsilon(c) + \sum_{n > L_\epsilon(c)} \mathsf{P}_A(T^R > n). \tag{110}$$

For any $n \in \mathbb{N}$, by the definition of T^R in (13) it follows that on the event $\{T^R > n\}$ there are $i \in A$ and $j \notin A$ such that $\Lambda_{ij}^R(n) < c$ and, as a result,

$$\mathsf{P}_A(T^R > n) \leq \sum_{i\in A, j\notin A} \mathsf{P}_A(\Lambda_{ij}^R(n) < c).$$

Moreover, for any $n > L_\epsilon(c)$ and $i \in A, j \notin A$,

$$c < n\left((c_i^*(A)-\epsilon)I_i + (c_j^*(A)-\epsilon)J_j - \epsilon\right), \tag{111}$$

and consequently for every $c > 0$ the series in (110) is bounded by

$$\sum_{i\in A, j\notin A} \sum_{n=1}^{\infty} \mathsf{P}_A\left(\frac{\Lambda_{ij}^R(n)}{n} < (c_i^*(A)-\epsilon)I_i + (c_j^*(A)-\epsilon)J_j - \epsilon \right). \tag{112}$$

By the assumption of the theorem and an application of Lemma A.2(iii) with ρ_i equal to $c_i^*(A) - \epsilon$ (resp. 0) when $x_A > 0$ (resp. $x_A = 0$) and ρ_j equal to $c_j^*(A) - \epsilon$ (resp. 0) when $y_A > 0$ (resp. $y_A = 0$), it follows that the series

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

in (112) converges. Thus, letting first $c \to \infty$ and then $\epsilon \to 0$ in (110) proves that, as $c \to \infty$,

$$\mathsf{E}_A[T^R] \lesssim \max_{i \in A, j \notin A} \frac{c}{c_i^*(A) I_i + c_j^*(A)\, J_j}. \tag{113}$$

In view of (46) and the selection of threshold c according to (19), this proves (108).

We next consider the case $\ell < u$, where $\alpha, \beta \to 0$ so that (7) holds for some $r \in (0, \infty)$. We prove the result when $\ell \le |A| < u$, as the proof when $\ell < |A| \le u$ is similar. Thus, in what follows, $\ell \le |A| < u$ and, as a result, $y_A > 0$ and $c_j^*(A) > 0$ for every $j \notin A$. By the universal asymptotic lower bounds (35) and (36) it follows that when either $|A| > \ell$ or $|A| = \ell = 0$, it suffices to show that

$$\mathsf{E}_A[T^R] \lesssim \frac{|\log \beta|}{y_A\, J_A^*}. \tag{114}$$

On the other hand, by the universal asymptotic lower bounds (37) and (39) it follows that when $|A| = \ell > 0$, it suffices to show that

$$\mathsf{E}_A[T^R] \lesssim \frac{|\log \beta|}{y_A J_A^*} \bigvee \frac{|\log \alpha|}{x_A I_A^* + y_A J_A^*}. \tag{115}$$

(When $r \le 1$, the maximum is attained strictly by the first term, when $r > 1$, $z_A < 1$ and $K > \hat{K}_A + z_A \check{K}_A$, the maximum is attained strictly by the second term, whereas in all other cases the two terms are equal to a first-order asymptotic approximation).

We start by proving (114) when $\ell < |A| < u$. In this case we also have $x_A > 0$, and consequently $c_i^*(A) > 0$ for every $i \in A$. Then, for $\epsilon > 0$ small enough and $a, b > 0$ we set

$$\begin{aligned} N_\epsilon(a,b) := \max_{j \notin A,\, i \in A} \Bigg\{ & \frac{a}{(c_j^*(A) - \epsilon) J_j - \epsilon}, \\ & \frac{b}{(c_i^*(A) - \epsilon) I_i - \epsilon} \Bigg\}, \end{aligned} \tag{116}$$

and observe that

$$\mathsf{E}_A[T^R] \le N_\epsilon(a,b) + \sum_{n > N_\epsilon(a,b)} \mathsf{P}_A(T^R > n). \tag{117}$$

By the definition of T^R in (17) it follows that, for any $n \in \mathbb{N}$, on the event $\{T^R > n\}$ there is either a $j \notin A$ such that $\Lambda_j^R(n) > -a$, or an $i \in A$ such that $\Lambda_i^R(n) < b$. As a result, by the union bound we obtain

$$\begin{aligned} \mathsf{P}_A(T^R > n) \le & \sum_{j \notin A} \mathsf{P}_A(-\Lambda_j^R(n) < a) \\ & + \sum_{i \in A} \mathsf{P}_A(\Lambda_i^R(n) < b). \end{aligned}$$

For any $n > N_\epsilon(a,b)$ and $i \in A$, $j \notin A$,

$$a < n\left((c_j^*(A) - \epsilon) J_j - \epsilon\right),$$

$$b < n\left((c_i^*(A) - \epsilon) I_i - \epsilon\right),$$

which implies that the series in (117) is bounded by

$$\begin{aligned} & \sum_{j \notin A} \sum_{n=1}^{\infty} \mathsf{P}_A \left(-\frac{1}{n} \Lambda_j^R(n) < (c_j^*(A) - \epsilon) J_j - \epsilon \right) \\ & + \sum_{i \in A} \sum_{n=1}^{\infty} \mathsf{P}_A \left(\frac{1}{n} \Lambda_i^R(n) < (c_i^*(A) - \epsilon) I_i - \epsilon \right). \end{aligned} \tag{118}$$

By the assumption of the theorem and an application of Lemma A.2(i) with $\rho_i = c_i^*(A) - \epsilon$ and of Lemma A.2(ii) with $\rho_j = c_j^*(A) - \epsilon$ it follows that (118) converges. Thus, letting first $a, b \to \infty$ and then $\epsilon \to 0$ in (117) proves that, as $a, b \to \infty$,

$$\mathsf{E}_A[T^R] \lesssim \max_{j \notin A,\, i \in A} \left\{ \frac{a}{c_j^*(A) J_j}, \frac{b}{c_i^*(A) I_i} \right\}.$$

In view of (46) and the selection of thresholds a, b according to (20), this implies that

$$\mathsf{E}_A[T^R] \lesssim \frac{|\log \beta|}{y_A\, J_A^*} \sim \frac{|\log \alpha|}{x_A\, I_A^*}, \tag{119}$$

and proves (114).

The proof when $|A| = \ell = 0$, in which case $x_A = 0$, is similar, with the difference that we use

$$N_\epsilon(a) := \max_{j \notin A} \left\{ \frac{a}{(c_j^*(A) - \epsilon) J_j - \epsilon} \right\} \tag{120}$$

in the place of $N_\epsilon(a, c)$, and apply only Lemma A.2(i).

It remains to show that (115) holds when $|A| = \ell > 0$, in which case x_A is not always positive. We recall the definitions of $L_\epsilon(c)$ and $N_\epsilon(a)$ in (109) and (120) and observe that for any $\epsilon > 0$ small enough and $a, c > 0$ we have

$$\begin{aligned} \mathsf{E}_A[T^R] \le & L_\epsilon(c) \vee N_\epsilon(a) \\ & + \sum_{n > L_\epsilon(c) \vee N_\epsilon(a)} \mathsf{P}_A(T^R > n). \end{aligned} \tag{121}$$

By the definition of T^R in (17) it follows that, for any $n \in \mathbb{N}$, on the event $\{T^R > n\}$ there are either $i \in A$ and $j \notin A$ such that $\Lambda_{ij}^R(n) < c$ or $j \notin A$ such that $\Lambda_j^R(n) > -a$, and as a result

$$\begin{aligned} \mathsf{P}_A(T^R > n) \le & \sum_{i \in A, j \notin A} \mathsf{P}_A(\Lambda_{ij}^R(n) < c) \\ & + \sum_{j \notin A} \mathsf{P}_A(\Lambda_j^R(n) > -a). \end{aligned}$$

Following similar steps as in the previous cases, applying in particular Lemma A.2(ii) with $\rho_j = c_j^*(A) - \epsilon$ and Lemma A.2(iii) with $\rho_j = c_j^*(A) - \epsilon$ and ρ_i equal to

$c_i^*(A) - \epsilon$ (resp. 0) when $x_A > 0$ (resp. $x_A = 0$), we conclude that as $a, c \to \infty$

$$\mathsf{E}_A[T^R] \lesssim \max_{i \in A, j \notin A} \left\{ \frac{a}{c_j^*(A) J_j} \bigvee \frac{c}{c_i^*(A) I_i + c_j^*(A) J_j} \right\}.$$

In view of (46) and the selection of thresholds a, c according to (20), this proves (115). ∎

Proof of Theorem 5.3: Fix $A \in \mathcal{P}_{\ell,u}$ and a probabilistic sampling rule R that satisfies (51). Since $c_i^*(A) > 0$ for every i in A (resp. A^c) when $x_A > 0$ (resp. $y_A > 0$), $x_A \vee y_A > 0$, $x_A > 0$ when $|A| > \ell$, and $y_A > 0$ when $|A| < u$, the exponentially consistency of R under P_A follows by an application of Theorem 4.1. To establish its asymptotic optimality, by Theorem 5.2 it follows that it suffices to show that $\mathsf{P}_A(\pi_i^R(n) < \rho)$ is an exponentially decaying sequence for every $\rho \in (0, c_i^*(A))$ and $i \in [M]$ such that $c_i^*(A) > 0$. Fix such i and ρ. Then, there is an $\epsilon > 0$ such that

$$\rho + \epsilon < c_i^*(A). \tag{122}$$

By the definition of a probabilistic rule (recall (22)), $R(n+1)$ is conditionally independent of $\mathcal{F}_n^R$ given Δ_n^R and its conditional distribution, q^R, does not depend on n. Thus, by [39, Prop. 6.13] there is a measurable function $h : \mathcal{P}_{\ell,u} \times [0,1] \to 2^{[M]}$, which does not depend on n, such that

$$R(n+1) = h\left(\Delta_n^R, Z_0(n)\right), \quad n \in \mathbb{N},$$

where $\{Z_0(n), n \in \mathbb{N}\}$ is a sequence of iid random variables, uniformly distributed in $(0,1)$. Consequently, there is a measurable function $h_i : \mathcal{P}_{\ell,u} \times [0,1] \to \{0,1\}$ such that

$$R_i(n+1) = h_i\left(\Delta_n^R, Z_0(n)\right), \quad n \in \mathbb{N}. \tag{123}$$

Then, for every $n \in \mathbb{N}$ we have

$$\begin{aligned} &\{\pi_i^R(n) < \rho\} \\ &= \Bigg\{ \sum_{m=1}^{n} h_i\left(A, Z_0(m-1)\right) \\ &\quad + \sum_{m=1}^{n} (R_i(m) - h_i(A, Z_0(m-1)) < n(\rho+\epsilon) - n\epsilon \Bigg\} \end{aligned}$$

and as a result $\mathsf{P}_A(\pi_i^R(n) < \rho)$ is upper bounded by

$$\begin{aligned} &\mathsf{P}_A\left(\sum_{m=1}^{n} h_i(A, Z_0(m-1)) < n(\rho+\epsilon)\right) \\ &+\mathsf{P}_A\left(\sum_{m=1}^{n} (R_i(m) - h_i(A, Z_0(m-1))) < -n\epsilon\right) \end{aligned} \tag{124}$$

From (23) and (123) it follows that $\{h_i(A, Z_0(n-1)), n \in \mathbb{N}\}$ is a sequence of iid Bernoulli random variables with parameter $c_i^R(A)$, whereas by (51) and (122) it follows that $\rho + \epsilon < c_i^R(A)$. Therefore, by the Chernoff bound we conclude that the first term in the upper bound in (124) is exponentially decaying. The second term is bounded as follows

$$\begin{aligned} &\mathsf{P}_A\left(\sum_{m=1}^{n} (R_i(m) - h_i(A, Z_0(m-1))) < -n\epsilon\right) \\ &\le \mathsf{P}_A\left(\sum_{m=1}^{n} |R_i(m) - h_i(A, Z_0(m-1))| > n\epsilon\right) \\ &\le \mathsf{P}_A\left(\sigma_A^R > n\right) + \mathsf{P}_A\left(\sigma_A^R > n\epsilon\right), \end{aligned} \tag{125}$$

where the first inequality follows from the triangle inequality and the second by an application of the total probability rule on the event $\{\sigma_A^R \le n\}$, in view of the fact that

$$\sigma_A^R \le n \;\Rightarrow\; \sum_{m=1}^{n} |R_i(m) - h_i(A, Z_0(m-1))| \le \sigma_A^R.$$

By the exponentially consistency of R, the upper bound in (125) is exponentially decaying, which means that the second term in the upper bound in (124) is also exponentially decaying, and this completes the proof. ∎

References

[1] J. Stiles and T. Jernigan, "The basics of brain development," *Neuropsychology review*, vol. 20, pp. 327–48, 11 2010.

[2] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235 – 255, 2002. [Online]. Available: https://doi.org/10.1214/ss/1042727940

[3] S. K. De and M. Baron, "Sequential bonferroni methods for multiple hypothesis testing with strong control of family-wise error rates i and ii," *Sequential Analysis*, vol. 31, no. 2, pp. 238–262, 2012.

[4] ——, "Step-up and step-down methods for testing multiple hypotheses in sequential experiments," *Journal of Statistical Planning and Inference*, vol. 142, no. 7, pp. 2059–2070, 2012.

[5] J. Bartroff and J. Song, "Sequential tests of multiple hypotheses controlling type i and ii familywise error rates," *Journal of statistical planning and inference*, vol. 153, pp. 100–114, 2014.

[6] Y. Song and G. Fellouris, "Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals," *Electronic Journal of Statistics*, vol. 11, 03 2016.

[7] ——, "Sequential multiple testing with generalized error control: An asymptotic optimality theory," *The Annals of Statistics*, vol. 47, no. 3, pp. 1776 – 1803, 2019. [Online]. Available: https://doi.org/10.1214/18-AOS1737

[8] J. Bartroff and J. Song, "Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates," *Sequential Analysis*, vol. 39, no. 1, pp. 65–91, 2020. [Online]. Available: https://doi.org/10.1080/07474946.2020.1726686

[9] K. S. Zigangirov, "On a problem in optimal scanning," *Theory of Probability & Its Applications*, vol. 11, no. 2, pp. 294–298, 1966. [Online]. Available: https://doi.org/10.1137/1111025

[10] E. M. Klimko and J. Yackel, "Optimal search strategies for wienér processes," *Stochastic Processes and their Applications*, vol. 3, pp. 19–33, 1975.

[11] V. Dragalin, "A simple and effective scanning rule for a multi-channel system," *Metrika*, vol. 43, pp. 165–182, 02 1996.

[12] K. Cohen and Q. Zhao, "Asymptotically optimal anomaly detection via sequential testing," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2929–2941, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3177142, IEEE Transactions on Information Theory

[13] B. Huang, K. Cohen, and Q. Zhao, "Active anomaly detection in heterogeneous processes," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2284–2301, 2019.

[14] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 831–852, 2018.

[15] A. Gurevich, K. Cohen, and Q. Zhao, "Sequential anomaly detection under a nonlinear system cost," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3689–3703, 2019.

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

[17] B. Hemo, T. Gafni, K. Cohen, and Q. Zhao, "Searching for anomalies over composite hypotheses," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1181–1196, 2020.

[18] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, vol. 30, no. 3, pp. 755–770, 09 1959. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177706205

[19] A. E. Albert, "The Sequential Design of Experiments for Infinitely Many States of Nature," *The Annals of Mathematical Statistics*, vol. 32, no. 3, pp. 774 – 799, 1961. [Online]. Available: https://doi.org/10.1214/aoms/1177704973

[20] S. A. Bessler, "Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments, part i theory." Department of Statistics, Stanford University, Technical Report 55, 1960.

[21] ——, "Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments, part ii applications." Department of Statistics, Stanford University, Technical Report 56, 1960.

[22] J. Kiefer and J. Sacks, "Asymptotically Optimum Sequential Inference and Design," *The Annals of Mathematical Statistics*, vol. 34, no. 3, pp. 705 – 750, 1963. [Online]. Available: https://doi.org/10.1214/aoms/1177704000

[23] R. Keener, "Second Order Efficiency in the Sequential Design of Experiments," *The Annals of Statistics*, vol. 12, no. 2, pp. 510 – 532, 1984. [Online]. Available: https://doi.org/10.1214/aos/1176346503

[24] S. P. Lalley and G. Lorden, "A Control Problem Arising in the Sequential Design of Experiments," *The Annals of Probability*, vol. 14, no. 1, pp. 136 – 172, 1986. [Online]. Available: https://doi.org/10.1214/aop/1176992620

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

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

[27] S. Nitinawarat and V. V. Veeravalli, "Controlled sensing for sequential multihypothesis testing with controlled markovian observations and non-uniform control cost," *Sequential Analysis*, vol. 34, no. 1, pp. 1–24, 2015. [Online]. Available: https://doi.org/10.1080/07474946.2014.961864

[28] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, "Sequential controlled sensing for composite multihypothesis testing," *Sequential Analysis*, vol. 40, no. 2, pp. 259–289, 2021. [Online]. Available: https://doi.org/10.1080/07474946.2021.1912525

[29] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*. PMLR, 2016, pp. 998–1027.

[30] S. I. Resnick, *Adventures in stochastic processes*. Springer Science & Business Media, 1992.

[31] G. Hommel and T. Hoffmann, "Controlled uncertainty," in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*. Springer, 1988, pp. 154–161.

[32] E. L. Lehmann and J. P. Romano, "Generalizations of the familywise error rate," *Ann. Statist.*, vol. 33, no. 3, pp. 1138–1154, 06 2005. [Online]. Available: http://dx.doi.org/10.1214/009053605000000084

[33] A. Tsopelakos and G. Fellouris, "Sequential anomaly detection with observation control under a generalized error metric," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1165–1170.

[34] J. Heydari, A. Tajer, and H. V. Poor, "Quickest linear search over correlated sequences," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5786–5808, 2016.

[35] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013. [Online]. Available: http://dx.doi.org/10.1561/0100000064

[36] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[37] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2015.

[38] Y. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, ser. Springer Texts in Statistics. Springer New York, 2012. [Online]. Available: https://books.google.com/books?id=213dBwAAQBAJ

[39] O. Kallenberg, *Foundations of Modern Probability*, ser. Probability and Its Applications. Springer New York, 2002. [Online]. Available: https://books.google.com/books?id=TBgFslMy8V4C

PLACE PHOTO HERE

Aristomenis Tsopelakos received his Diploma in Electrical and Computer Engineering from National Technical University of Athens in 2014, his Masters in Electrical and Computer Engineering in 2016 and his Masters in Mathematics in 2018 both from University of Illinois Urbana-Champaign. His research interests include detection and estimation theory, sequential design for multiple hypothesis testing, control of stochastic systems. He is currently a PhD student at Coordinated Science Laboratory and Electrical and Computer Engineering department at University of Illinois Urbana-Champaign.

PLACE PHOTO HERE

Georgios Fellouris received the Diploma in Applied Mathematics in 2004 from the National Technical University of Athens, and the Ph.D. degree in Statistics from Columbia University in 2010. He was a postdoctoral research associate at the Department of Mathematics in the University of Southern California during 2010-2013. He joined the University of Illinois, Urbana-Champaign in 2013, where he is currently an Associate Professor in the Department of Statistics and affiliated with the Electrical and Computer Engineering Department and the Coordinated Science Laboratory. His research interests include sequential detection and estimation, quickest change detection, adaptive experimental design.