The Sample Complexity of Search over Multiple Populations

Matthew L. Malloy
Electrical and Computer Engineering
University of Wisconsin-Madison
Email: mmalloy@wisc.edu

Gongguo Tang
Electrical and Computer Engineering
University of Wisconsin-Madison
Email: gtang5@wisc.edu

Robert D. Nowak
Electrical and Computer Engineering
University of Wisconsin-Madison
Email: nowak@ece.wisc.edu

***Abstract*—This paper studies the sample complexity of searching over multiple populations. We consider a large number of populations, each corresponding to either distribution P_0 or P_1. The goal of the search problem studied here is to find one population corresponding to distribution P_1 with as few samples as possible. The main contribution is to precisely quantify the number of samples needed to correctly find one such population. We consider two general approaches: non-adaptive sampling methods, which sample each population a predetermined number of times until a population following P_1 is found, and adaptive sampling methods, which employ sequential sampling schemes for each population. We first derive a lower bound on the number of samples required by any sampling scheme. We then consider an adaptive procedure consisting of a series of sequential probability ratio tests, and show it comes within a small constant factor of the lower bound. We give explicit expressions for this constant when samples of the populations follow Gaussian and Bernoulli distributions. An alternative adaptive scheme is discussed which does not require full knowledge of P_1, and outperforms non-adaptive schemes. For comparison, a lower bound on the sampling requirements of any non-adaptive scheme is presented.**

***Index Terms*—Quickest search, rare events, SPRT, CUSUM procedure, sparse recovery, sequential analysis, sequential thresholding, biased coin, spectrum sensing, multi-armed bandit.**

I. Introduction

This paper studies the sample complexity of finding a population corresponding to some distribution P_1 among a large number of populations corresponding to either distribution P_0 or P_1. More specifically, let $i = 1, 2, \ldots$ index the populations. Samples of each population follow one of two distributions, indicated by a binary label X_i: if $X_i = 0$, then samples of population i follow distribution P_0, if $X_i = 1$, then samples follow distribution P_1. We assume that $X_1, X_2, \ldots$ are independently and identically distributed (i.i.d.) Bernoulli random variables with $\mathbb{P}(X_i = 0) = 1-\pi$ and $\mathbb{P}(X_i = 1) = \pi$. Distribution P_1 is termed the *atypical* distribution, which corresponds to *atypical* populations, and the probability π quantifies the occurrence of such populations. The goal of the search problem studied here is to find an atypical population with as few samples as possible.

In this search problem, populations are sampled a (deterministic or random) number of times, in sequence, until an atypical population is found. The total number of samples needed is a function of the sampling strategy, the distributions, the required reliability, and π. To build intuition, consider the following. As the occurrence of the atypical populations becomes infrequent, (i.e. as $\pi \to 0$), the number of samples required to find one such population must, of course, increase. If P_0 and P_1 are extremely different (e.g., non-overlapping supports), then a search procedure could simply proceed by taking one sample of each population until an atypical population was found. The procedure would identify an atypical population with, on average, π^{-1} samples. More generally, when the two distributions are more difficult to distinguish, as is the concern of this paper, we must take multiple samples of some populations. As the required reliability of the search increases, a procedure must also take more samples to confirm, with increasing certainty, that an atypical population has been found.

The main contribution of this work is to precisely quantify the number of samples needed to correctly find one atypical population. Specifically, we provide tight bounds on the expected number of samples required to find a population corresponding to P_1 with a specified level of certainty. We pay additional attention to this sample complexity as π becomes small (and the occurrence of the atypical populations becomes *rare*). We consider two general approaches to find an atypical population, both of which sample populations in sequence. *Non-adaptive* procedures sample each population a predetermined number of times, make a decision, and if the null hypothesis is accepted then move on to the next population. *Adaptive* methods, in contrast, enjoy the flexibility to sample each population sequentially, and thus, the decision to continue sampling a particular population can be based on prior samples.

The developments in this paper proceed as follows. First, using techniques from sequential analysis, we derive a lower bound on the expected number of samples needed to reliably identify an atypical population. To preview the results, the lower bound implies that any procedure (adaptive or non-adaptive) is unreliable if it uses fewer than $\pi^{-1}D(P_0||P_1)^{-1}$ samples on average, where $D(P_0||P_1)$ is the Kullback-Leibler divergence. We then prove this is tight by showing that a series of sequential probability ratio tests (which we abbreviate as an S-SPRT) succeeds with high probability if the total number

This work was presented in part at the Conference on Information Sciences and Systems (CISS) in Princeton, USA [1].

of samples is within a constant factor of the lower bound, provided a minor constraint on the log-likelihood statistic is satisfied (which holds for bounded distributions, Gaussian, exponential, among others). We give explicit expressions for this constant in the Gaussian and Bernoulli cases. In the Bernoulli case, the bound derived by instantiating our general results produces the tightest known bound. In many real world problems, insufficient knowledge of the distributions P_0 and P_1 makes implementing an S-SPRT impractical. To address this shortcoming, we propose a more practical adaptive procedure known as sequential thresholding, which doesn't require precise knowledge of P_1, and is particularly well suited for problems in which occurrence of an atypical population is rare. We show sequential thresholding is near-optimal when $\pi \to 0$. Both the S-SPRT procedure and sequential thresholding are shown to be robust to imperfect knowledge of π. Lastly, we show that non-adaptive procedures require at least $\pi^{-1} D(P_1||P_0)^{-1} \log \pi^{-1}$ samples to reliably find an atypical population, a factor of $\log \pi^{-1}$ more samples when compared to adaptive methods.

A. Motivating Applications

Finding an atypical population arises in many relevant problems in science and engineering. One of the main motivations for our work is the problem of spectrum sensing in cognitive radio. In cognitive radio applications, one is interested in finding a vacant radio channel among a potentially large number of occupied channels. Only once a vacant channel is identified can the cognitive device transmit, and thus, identifying a vacant channel as quickly as possible is of great interest. A number of works have looked at various adaptive methods for spectrum sensing in similar contexts, including [2]–[4].

Another captivating example is the *Search for Extraterrestrial Intelligence* (SETI) project. Researchers at the SETI institute use large antenna arrays to sense for narrowband electromagnetic energy from distant star systems, with the hopes of finding extraterrestrial intelligence with technology similar to ours. The search space consists of a virtually unlimited number of stars, over 100 billion in the Milky Way alone, each with 9 million potential "frequencies" in which to sense for narrow band energy. The prior probability of extraterrestrial transmission is indeed very small (SETI has yet to make a "contact"), and thus occurrence of atypical populations is rare. Roughly speaking, SETI employs a variable sample size search procedure that repeatedly tests energy levels against a threshold up to five times [5], [6]. If any of the measurements are below the threshold, the procedure immediately passes to the next frequency/star. This procedure is closely related to *sequential thresholding* [7]. Sequential thresholding results in substantial gains over fixed sample size procedures and, unlike the SPRT, it can be implemented without perfect knowledge of P_1.

B. Related Work

The prior work most closely related to the problem investigated here is that by Lai, Poor, Xin, and Georgiadis [8], in which the authors also examine the problem of quickest search across multiple populations, but do not focus on quantifying the sample complexity. The authors show that the S-SPRT (also termed a CUSUM test) optimizes a linear combination of the expected number of samples and the error probability. Complementary to this, our contributions include providing tight lower bounds on the expected number of samples required to achieve a desired probability of error, and then showing the sample complexity of the S-SPRT comes within a small constant of this bound. This quantifies how the number of samples required to find an atypical population depends on the distributions P_0 and P_1 and the probability π, which was not explicitly investigated in [8]. As a by-product, this proves the optimality of the S-SPRT.

An instance of the quickest search problem was also studied recently in [9], where the authors investigate the problem of finding a biased coin with the fewest flips. Our more general results are derived using different techniques, and cover this case with P_0 and P_1 as Bernoulli distributions. In [9], the authors present a bound on the expected number of flips needed to find a biased coin. The bound derived from instantiating our more general theory (see example 2 and Corollary 8) is a minimum of 32 times tighter than the bound in [9].

Also closely related is the problem of sparse signal support recovery from point-wise observations [7], [10], [11], classical work in optimal scanning theory [12], [13], and work on pure exploration in multi-armed bandit problems [14], [15]. The sparse signal recovery problems differ in that the total number of populations is finite, and the objective is to recover all (or most) populations following P_1, as opposed to finding a single population and terminating the procedure. Traditional multi-armed bandit problems differ in that no knowledge of the distributions of the arms is assumed.

II. Problem Setup

Consider an infinite number of populations indexed by $i = 1, 2, ...$ For each population i, samples of that population are distributed either

$$\begin{aligned} Y_{i,j} &\overset{iid}{\sim} P_0 \text{ if } X_i = 0 \quad \text{or} \\ Y_{i,j} &\overset{iid}{\sim} P_1 \text{ if } X_i = 1 \end{aligned}$$

where P_0 and P_1 are probability measures supported on $\mathcal{Y}$, j indexes multiple i.i.d. samples of a particular population, and X_i is a binary label. The goal is to find a population i such that $X_i = 1$ as quickly and reliably as possible. The prior probability of a particular population i following P_1 or P_0 is i.i.d., and denoted

$$\begin{aligned} \mathbb{P}(X_i = 1) &= \pi \\ \mathbb{P}(X_i = 0) &= 1 - \pi \end{aligned}$$

where we assume $\pi \leq 1/2$ without loss of generality.

Algorithm 1 Search for an atypical population

initialize: $i = 1$, $j = 1$
while atypical population not found **do**
  sample $Y_{i,j}$
  either
    1) re-sample population i: $j = j + 1$
    2) move to next population: $i = i + 1$, $j = 1$
    3) terminate: $\hat{X}_i = 1$
end while
output: $I = i$

Also without loss of generality, a testing procedure starts at population $i = 1$ and takes one sample. The procedure then decides to either *1)* take an additional sample of $i = 1$, or *2)* estimate population $i = 1$ as following distribution P_0 (deciding $\hat{X}_1 = 0$) and move to index 2, or *3)*, terminate, declaring population $i = 1$ as following distribution P_1 (deciding $\hat{X}_1 = 1$). Provided the procedure doesn't terminate, it continues in this fashion, taking one of three actions after each sample is taken. As in [8], the procedure does not revisit populations (which is well justified as each population is independent of all others).

The performance of any testing procedure is characterized by two metrics: *1)* the expected number of samples required for the procedure to terminate, denoted $\mathbb{E}[N]$, and *2)* the probability the procedure returns an index not corresponding to a population following P_1. We denote this probability as

$$P_e := \mathbb{P}\left(I \in \{i : X_i = 0\}\right)$$

where I is a random variable representing the index on which the procedure terminates.

Imagine that the procedure is currently sampling index i. For a given sampling procedure, if $X_i = 1$, the probability the procedure passes to index $i+1$ without terminating is denoted β, and the probability the procedure correctly declares $\hat{X}_i = 1$ is $1 - \beta$. Likewise, for any i such that $X_i = 0$, the procedure falsely declares $\hat{X}_i = 1$ with probability α, and continues to index $i + 1$ with probability $1 - \alpha$. In other words, provided the procedure arrives at population i,

$$\begin{aligned} \beta &= \mathbb{P}(\hat{X}_i = 0 | X_i = 1) \\ \alpha &= \mathbb{P}(\hat{X}_i = 1 | X_i = 0). \end{aligned}$$

In essence, the procedure consists of a number of simple binary hypothesis tests, each with false positive probability α and false negative probability β.

The following recursive relationships will be central to our performance analysis. Let N_i be the (random) number of samples taken of population i, and $N = \sum_{i=1}^{\infty} N_i$ be the total number of samples taken by the procedure. We can write the expected number of samples as

$$\begin{aligned} \mathbb{E}[N] &= \mathbb{E}[N_1] \quad + \\ &\quad \mathbb{E}\left[N_2 + N_3 + ... \,\middle|\, \hat{X}_1 = 0\right]((1-\pi)(1-\alpha) + \pi\beta) \end{aligned} \tag{1}$$

where $(1 - \pi)(1 - \alpha) + \pi\beta$ is the probability the procedure arrives at the second index. The expected number of samples used from the second index onwards, given that the procedure arrives at the second index (without declaring $I = 1$), is simply equal to the total number of samples: $\mathbb{E}[N_2 + N_3 + ... \,|\hat{X}_1 = 0] = \mathbb{E}[N]$. Rearranging terms in (1) gives the following relationship

$$\mathbb{E}[N] = \frac{\mathbb{E}[N_1]}{\alpha(1-\pi) + \pi(1-\beta)}. \tag{2}$$

For simplicity of notation, denote the expected number of measurements conditioned on the binary label as

$$E_1 = \mathbb{E}[N_1 | X_1 = 1] \qquad E_0 = \mathbb{E}[N_1 | X_1 = 0]$$

and thus,

$$\mathbb{E}[N] = \frac{\pi E_1 + (1-\pi)E_0}{\alpha(1-\pi) + \pi(1-\beta)}. \tag{3}$$

In the same manner we arrive at the following expression for the probability of error:

$$\begin{aligned} P_e &= \frac{\alpha(1-\pi)}{\alpha(1-\pi) + \pi(1-\beta)} \\ &= \frac{1}{1 + \frac{\pi(1-\beta)}{\alpha(1-\pi)}}. \end{aligned} \tag{4}$$

From this expression we see that if

$$\frac{\alpha(1-\pi)}{\pi(1-\beta)} \geq \delta$$

for some $\delta > 0$, then $P_e \geq \frac{\delta}{1+\delta}$, and P_e is greater than or equal to some positive constant.

Lastly, the bounds derived throughout often depend on explicit constants, in particular the Kullback-Leibler divergence between the distributions, which is defined in the usual manner:

$$D(P_0 || P_1) = \mathbb{E}_0\left[\log \frac{P_0(Y)}{P_1(Y)}\right].$$

Other constants are denoted by C_1, C_1', etc., and represent distinct numerical constants.

III. Lower bound for any procedure

We begin with a lower bound on the number of samples required by any procedure to find a population following distribution P_1. Before stating the main theorem of the section, we present a number of corollaries of Theorem 5 aimed at highlighting the explicit relationship between the problem parameters.

Corollary 1. *Any procedure with*

$$P_e \leq \frac{\delta}{1+\delta}$$

also has

$$\mathbb{E}[N] \geq \frac{1}{D(P_0||P_1)}\left(\frac{1}{12\,\pi} + \frac{1}{3}\,\log\left(\frac{1}{2\pi\delta}\right) - 1\right) \tag{5}$$

for any $\delta \leq 1/2$. Here, we assume $D(P_0||P_1) = D(P_1||P_0)$ for simplicity of presentation.

Proof of Corollary 1 follows immediately from Theorem 5, as $\pi \leq 1/2$ and $\delta \leq 1/2$.

Corollary 1 provides a particularly intuitive way to quantify the number of samples required for the quickest search problem. The first term in (5), which has a $1/\pi$ dependence, can be interpreted as the minimum number of samples required to *find* a population following distribution P_1. The second term, which has a $\log \delta^{-1}$ dependence, is best interpreted as the minimum number of samples required to *confirm* that a population following P_1 has been found.

When the populations following distribution P_1 become rare (when π tends to zero), the second and third terms in (5) become small compared to the first term. This suggests the vast majority of samples are used to *find* a rare population, and a vanishing proportion are needed for *confirmation*. The corollaries below capture this effect. The leading constants are of particular importance, as we relate them to upper bounds in Sec. IV. In the following, consider P_e and $\mathbb{E}[N]$ as functions π, P_0, P_1, and some sampling procedure $\mathcal{A}$.

Corollary 2. Rare population. *Fix $\delta \in (0, 1/2]$. Then any procedure $\mathcal{A}$ that satisfies*

$$\limsup_{\pi \to 0} P_e \leq \frac{\delta}{1+\delta}$$

also has

$$\liminf_{\pi \to 0} \ \pi \, \mathbb{E}[N] \quad \geq \quad \frac{(1-\delta)^2}{(1+\delta)} \max\left(\frac{1}{D(P_0||P_1)}, 1\right).$$

The proof of Corollary 2 follows from Theorem 5 by noting both the second and third terms of (7) are overwhelmed as π becomes small. The lower bound in in Corollary 2 states that any procedure requiring fewer than order $1/\pi$ samples is unreliable, and is best interpreted in two regimes: *(1)* the high SNR regime, when $D(P_0||P_1) > 1$, and *(2)*, the low SNR regime, when $D(P_0||P_1) \leq 1$.

Corollary 3. High SNR. *When $D(P_0||P_1) > 1$, any procedure with $\lim_{\pi \to 0} P_e = 0$ also has $\lim_{\pi \to 0} \ \pi \, \mathbb{E}[N] \geq 1$.*

The proof follows from Corollary 2. Corollary 3 states that any procedure requiring fewer samples in expectation than π^{-1} also has probability of error bound away from zero. The bound becomes tight when the SNR becomes high – when $D(P_0||P_1)$ is sufficiently large, we expect to classify each population with one sample.

Corollary 4. Low SNR. *If $D(P_0||P_1) \leq 1$, any procedure with $\lim_{\pi \to 0} P_e = 0$ also has $\lim_{\pi \to 0} \pi \mathbb{E}[N] \geq 1/D(P_0||P_1)$.*

Again the proof follows from Corollary 2. The Corollary simply indicates the following: in the low SNR regime the sampling requirements are at best an additional factor of $D(P_0||P_1)^{-1}$ higher than when we can classify each distribution with one sample.

Next we state a general lower bound in the main theorem of the section.

Theorem 5. *Any procedure with*

$$P_e \leq \frac{\delta}{1+\delta}$$

also has

$$\mathbb{E}[N] \quad \geq \quad \frac{1-\pi}{\pi} \frac{(1-\delta)^2}{(1+\delta)} \max\left(1, \frac{1}{D(P_0||P_1)}\right) + \frac{\log\left(\frac{1}{2\pi\delta}\right)}{D(P_1||P_0)} \left(\frac{1 - \delta \frac{D(P_1||P_0)}{D(P_0||P_1)}}{1+\delta}\right) - \frac{1}{D(P_1||P_0)} \tag{6}$$

for any $\delta \in [0, 1/2]$.

Proof: Assume that $P_e \leq \frac{\delta}{1+\delta}$ and from (4) we have

$$\frac{\alpha(1-\pi)}{\pi(1-\beta)} \leq \delta. \tag{7}$$

From (2),

$$\begin{aligned} \mathbb{E}[N] &= \frac{\pi E_1 + (1-\pi)E_0}{\alpha(1-\pi) + \pi(1-\beta)} \geq \frac{\pi E_1 + (1-\pi)E_0}{(1+\delta)\pi(1-\beta)} \\ &= \frac{E_1}{(1+\delta)(1-\beta)} + \frac{(1-\pi)E_0}{(1+\delta)\pi(1-\beta)}. \end{aligned} \tag{8}$$

From standard sequential analysis techniques (see Theorem 2.29 of [16]) we have the following identities relating the expected number of measurements to α and β, which hold for *any* binary hypothesis testing procedure:

$$E_1 \quad \geq \quad \frac{\beta \log\left(\frac{\beta}{1-\alpha}\right) + (1-\beta)\log\left(\frac{1-\beta}{\alpha}\right)}{D(P_1||P_0)} \tag{9}$$

$$E_0 \quad \geq \quad \frac{\alpha \log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)}{D(P_0||P_1)}. \tag{10}$$

Rearranging (8),

$$\mathbb{E}[N] \quad \geq \quad \underbrace{\frac{\beta \log\left(\frac{\beta}{1-\alpha}\right)}{(1+\delta)(1-\beta)D(P_1||P_0)}}_{T_1} + \underbrace{\frac{\log\left(\frac{1-\beta}{\alpha}\right)}{(1+\delta)D(P_1||P_0)}}_{T_2} + \underbrace{\frac{(1-\pi)\left(\alpha \log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)\right)}{\pi(1+\delta)(1-\beta)D(P_0||P_1)}}_{T_3}.$$

We first bound T_1 as

$$T_1 \geq \frac{-1}{(1+\delta)D(P_1||P_0)} \geq \frac{-1}{D(P_1||P_0)} \tag{11}$$

since for all $\beta \in [0, 1]$,

$$\frac{\beta \log \frac{\beta}{1-\alpha}}{1-\beta} \geq \frac{\beta \log \beta}{1-\beta} \geq -1.$$

From (7),

$$T_2 \geq \frac{\log\left(\frac{1-\pi}{\pi\delta}\right)}{(1+\delta)D(P_1||P_0)}.$$

Next, differentiating T_3 with respect to α gives

$$\frac{d(T_3)}{d\alpha} = \frac{(1-\pi)\log\frac{\alpha\beta}{(1-\alpha)(1-\beta)}}{(1-\delta)\pi(1-\beta)D(P_0||P_1)}$$

showing that the expression is non-increasing in α over the set of α satisfying $\frac{\alpha}{1-\beta} \leq \frac{1-\alpha}{\beta}$. From (7), we are restricted to $\frac{\alpha}{1-\beta} \leq \frac{\delta\pi}{1-\pi}$ and thus, if $\frac{\delta\pi}{1-\pi} \leq \frac{1-\alpha}{\beta}$, then (10) is non-increasing in α. To show this, note that

$$\frac{\delta\pi}{1-\pi} \leq \delta \leq 1-\delta \leq 1 - \frac{\alpha(1-\pi)}{\pi(1-\beta)} \leq 1-\alpha \leq \frac{1-\alpha}{\beta}$$

since both $\delta \leq 1/2$ and $\pi \leq 1/2$. We can replace α in (10) with $\frac{\delta\pi(1-\beta)}{1-\pi}$. This gives

$$\begin{aligned} T_3 &\geq \frac{\delta\log\left(\frac{\delta\pi}{1-\pi}\right)}{(1+\delta)D(P_0||P_1)} + \\ &\quad \frac{(1-\pi)\left(1-\frac{\delta\pi(1-\beta)}{1-\pi}\right)\log\left(\frac{1}{\beta}-\frac{\delta\pi(1-\beta)}{\beta(1-\pi)}\right)}{\pi D(P_0||P_1)(1+\delta)(1-\beta)} \\ &\geq \frac{\delta\log\left(\frac{\delta\pi}{1-\pi}\right)}{(1+\delta)D(P_0||P_1)} + \\ &\quad \frac{(1-\pi)(1-\delta)\log\left(\frac{1}{\beta}-\frac{\delta(1-\beta)}{\beta}\right)}{\pi D(P_0||P_1)(1+\delta)(1-\beta)} \\ &\geq \frac{\delta\log\left(\frac{\delta\pi}{1-\pi}\right)}{(1+\delta)D(P_0||P_1)} + \frac{(1-\pi)(1-\delta)^2}{\pi D(P_0||P_1)(1+\delta)} \end{aligned}$$

where the first inequality follows from making the substitution for α and from (7), and the second inequality follows since $\pi/(1-\pi) \leq 1$ and $1-\beta \leq 1$, and the last inequality follows as

$$\frac{\log\left(\frac{1}{\beta}-\frac{\delta(1-\beta)}{\beta}\right)}{(1-\beta)} \geq 1-\delta \tag{12}$$

for all $\beta \in [0,1]$. To see the validity of (12), we note

$$\begin{aligned} \frac{\log\left(\frac{1}{\beta}-\frac{\delta(1-\beta)}{\beta}\right)}{(1-\beta)(1-\delta)} &= \frac{\log\left(1+\frac{(1-\delta)(1-\beta)}{\beta}\right)}{(1-\beta)(1-\delta)} \\ &\overset{(\star)}{\geq} \frac{\log\left(1+\frac{1-\beta}{\beta}\right)}{1-\beta} \\ &= \frac{\log(1/\beta)}{1-\beta} \\ &\geq 1. \end{aligned}$$

Here $(\star)$ follows by noting that $\log(1+x/\beta)/x$ is monotonically decreasing in x, and by setting $x = 1-\beta$.

We can also trivially bound E_0 by noting that $E_0 \geq 1$. This provides an additionally bound on T_3:

$$\begin{aligned} T_3 &\geq \frac{(1-\pi)}{\pi(1+\delta)(1-\beta)} \\ &\geq \frac{\delta\log\left(\frac{\delta\pi}{1-\pi}\right)}{(1+\delta)D(P_0||P_1)} + \frac{(1-\pi)(1-\delta)^2}{\pi(1+\delta)} \end{aligned} \tag{13}$$

since the first term in (13) is strictly negative.

Combining the bounds on T_1 and T_2, and the two bounds on T_3, and noting that $\delta\pi/(1-\pi) \leq 2\delta\pi$ gives

$$\begin{aligned} \mathbb{E}[N] \quad \geq \quad & \frac{1-\pi}{\pi}\frac{(1-\delta)^2}{(1+\delta)}\max\left(1, \frac{1}{D(P_0||P_1)}\right) + \\ & \frac{\log\left(\frac{1}{2\pi\delta}\right)}{D(P_1||P_0)}\left(\frac{1-\delta\frac{D(P_1||P_0)}{D(P_0||P_1)}}{1+\delta}\right) - \frac{1}{D(P_1||P_0)} \end{aligned}$$

completing the proof. ∎

IV. S-SPRT Procedure

The Sequential Probability Ratio Test (SPRT), optimal for simple binary hypothesis tests in terms of minimizing the expected number of samples for tests of a given power [17], can be applied to the problem studied here by implementing a series of SPRTs on the individual populations. For notational convenience, we refer to this procedure as the S-SPRT. This is equivalent in form to the CUSUM test studied in [8], which is traditionally applied to change point detection problems.

The S-SPRT operates as follows. Imagine the procedure has currently taken j samples of population i. The procedure continues to sample population i provided

$$\gamma_{\mathrm{L}} \ < \ \ell_{i,j} \ < \ \gamma_{\mathrm{U}}. \tag{14}$$

where $\ell_{i,j} := \prod_{k=1}^{j} \frac{P_1(Y_{i,k})}{P_0(Y_{i,k})}$ is the likelihood ratio statistic, and γ_{U} and γ_{L} are scalar upper and lower thresholds. In words, the procedure continues to sample population i provided the likelihood ratio comprised of samples of that population is between two scalar thresholds. The S-SPRT stops sampling population i after N_i samples, which is a random integer representing the smallest number of samples such that (14) no longer holds:

$$N_i := \min\left\{j : \ell_{i,j} \leq \gamma_{\mathrm{L}} \bigcup \ell_{i,j} \geq \gamma_{\mathrm{U}}\right\}.$$

When the likelihood ratio exceeds (or equals) γ_{U}, then $\hat{X}_i = 1$, and the S-SPRT terminates returning $I = i$. Conversely, if the likelihood ratio falls below (or equals) γ_{L}, then $\hat{X}_i = 0$, and the procedure moves to index $i+1$. The procedure is detailed in Algorithm 2.

Algorithm 2 Series of SPRTs Procedure (S-SPRT)

input: thresholds γ_{L}, γ_{U}, distributions P_0, P_1
initialize: $i = 1$, $j = 1$, $\ell = 1$
while $\ell \ < \ \gamma_{\mathrm{U}}$ **do**
  measure: $Y_{i,j}$
  compute: $\ell = \ell \cdot \frac{P_1(Y_{i,j})}{P_0(Y_{i,j})}$
  if $\ell \leq \gamma_L$ **then**
    $i = i+1$, $j = 1$, $\ell = 1$
  else
    $j = j+1$
  end if
end while
output: $I = i$

The S-SPRT procedure studied in [8] fixes the lower threshold in each individual SPRT at $\gamma_{\rm L} = 1$, which has a very intuitive interpretation. Since there are an infinite number of populations, anytime a sample suggests that a particular population doesn't follow P_1, moving to another population is best. While this approach is optimal [8], we use a strictly smaller threshold, as it results in a simpler derivation of the upper bound.

In the following theorem and corollary we assume a minor restriction on the tail distribution of the log-likelihood ratio test statistic, a notion studied in depth in [18]. Specifically, let $L = \log(P_1(Y)/P_0(Y))$ be the log-likelihood statistic. We require that

$$\max_{r\geq 0} \mathbb{E}\left[L - r | L \geq r\right] < \infty \tag{15}$$

and

$$\min_{r\geq 0} \mathbb{E}\left[L + r | L \leq -r\right] > -\infty. \tag{16}$$

This condition is satisfied when L follows any bounded distribution, Gaussian distributions, exponential distributions, among others. It is not satisfied by distributions with infinite variance or polynomial tails. A more thorough discussion of this restriction is studied in [18].

Corollary 6. Rare population. *Fix $\delta \in (0, 1/2]$. The S-SPRT with any $\gamma_{\rm L} \in (0,1)$ and $\gamma_{\rm U} = \frac{1-\pi}{\pi\delta}$ satisfies $P_e \leq \frac{\delta}{1+\delta}$ and*

$$\lim_{\pi\to 0} \pi\, \mathbb{E}[N] \leq \frac{C_1}{D(P_0||P_1)}$$

for some constant C_1 independent of π and δ.

The proof of Corollary 6 is an immediate consequence of Theorem 7. Note that $\gamma_{\rm U} > 1$, since we assume that $\pi \leq 1/2$. As the atypical populations become rare, sampling is dominated by *finding* an atypical population, which is order π^{-1}. The constant factor of $C_1/D(P_0||P_1)$ is the multiplicative increase in the number of samples required when the problem becomes noisy. C_1 can be explicitly calculated in a number of scenarios (see Examples 1 and 2).

Theorem 7. *The S-SPRT with $\gamma_{\rm L} \in (0,1)$ and $\gamma_{\rm U} = \frac{1-\pi}{\pi\delta}$, $\delta \in [0, 1/2]$ satisfies*

$$P_e \leq \frac{\delta}{1+\delta}$$

and

$$\mathbb{E}[N] \leq \frac{C_1}{\pi D(P_0||P_1)} + \frac{\log\frac{1}{\pi\delta}}{D(P_1||P_0)} + \frac{C_2}{D(P_1||P_0)} \tag{17}$$

for some constants C_1 and C_2 independent of π and δ.

Proof: The proof is based on techniques used for analysis of the SPRT. From [16], the false positive and false negative events are related to the thresholds as:

$$\alpha \leq \gamma_{\rm U}^{-1}(1-\beta) \leq \gamma_{\rm U}^{-1} = \frac{\pi\delta}{1-\pi} \tag{18}$$

$$\beta \leq \gamma_{\rm L}(1-\alpha) \leq \gamma_{\rm L}. \tag{19}$$

From (4) the probability the procedure terminates in error returning a population following P_0 is

$$P_e \;\leq\; \frac{1}{1+\frac{\pi}{\gamma_{\rm U}^{-1}(1-\pi)}} = \frac{\delta}{1+\delta}. \tag{20}$$

To show the second part of the theorem, first define the log-likelihood ratio as

$$L_i^{(j)} = \sum_{k=1}^{j} \log\frac{P_1(Y_{i,k})}{P_0(Y_{i,k})}. \tag{21}$$

By Wald's identity [16],

$$E_0 \;=\; \frac{-\mathbb{E}_0\left[L_i^{(N_i)}\right]}{D(P_0||P_1)} \;=\; \frac{(1-\alpha)\mathbb{E}_0\left[-L_i^{(N_i)}\middle|\hat{X}=0\right] + \alpha\mathbb{E}_0\left[-L_i^{(N_i)}\middle|\hat{X}=1\right]}{D(P_0||P_1)}.$$

The expected value of the log-likelihood ratio after N_i samples (i.e, when the procedure stops sampling index i) is often approximated by the stopping boundaries themselves (see [16]). In our case, it is sufficient to show the value of the likelihood ratio when the procedure terminates or moves to the next index can be bound by a constant independent of π and δ. From [17], equations 4.9 and 4.10, for $C_1' \geq 0$,

$$\mathbb{E}_0\left[L_i^{(N_i)}\middle|\hat{X}=0\right] \geq \log\gamma_{\rm L} - C_1' \tag{22}$$

and

$$\mathbb{E}_0\left[L_i^{(N_i)}\middle|\hat{X}=1\right] \leq \log\gamma_{\rm U} + C_1' \tag{23}$$

where C_1' is any constant that satisfies both

$$C_1' \leq \max_{r\geq 0} \mathbb{E}_0\left[L^{(1)} - r \middle| L^{(1)} \geq r\right]$$

and

$$C_1' \leq \max_{r\geq 0} \mathbb{E}_0\left[-(L^{(1)} + r)\middle| L^{(1)} \leq -r\right].$$

C_1' depends only on the distribution of $L^{(1)}$, and is trivially independent of $\gamma_{\rm L}$ and $\gamma_{\rm U}$. Under the assumptions of (15) and (16), the constants are finite. C_1' can be explicitly calculated for a variety of problems (see Examples 1 and 2, and [19], p.145, and [18]). C_1' is a bound on the *overshoot* in the log-likelihood ratio when it falls outside γ_U or γ_L. We have

$$\begin{aligned} E_0 \;&\leq\; \frac{(1-\alpha)\left(C_1' + \log\gamma_{\rm L}^{-1}\right) + \alpha\left(-C_1' + \log\gamma_{\rm U}^{-1}\right)}{D(P_0||P_1)} \\ &\leq\; \frac{(1-\alpha)(C_1' + \log\gamma_{\rm L}^{-1})}{D(P_0||P_1)} \end{aligned}$$

where the second inequality follows as $\gamma_{\rm U} \geq 1$. Likewise,

$$\begin{aligned} E_1 \;&\leq\; \frac{(1-\beta)(C_2' + \log\gamma_{\rm U}) + \beta(-C_2' + \log\gamma_{\rm L})}{D(P_1||P_0)} \\ &\leq\; \frac{(1-\beta)(C_2' + \log\gamma_{\rm U})}{D(P_1||P_0)} \end{aligned}$$

for some constant $C_2' \geq 0$ which represents the overshoot of the log-likelihood ratio given $X_i = 0$. Combining these with (3) bounds the expected number of samples:

$$\begin{aligned}\mathbb{E}[N] &\leq \frac{\pi\frac{(1-\beta)(C_2'+\log\gamma_{\mathrm{U}})}{D(P_1||P_0)} + (1-\pi)\frac{(1-\alpha)(C_1'+\log\gamma_{\mathrm{L}}^{-1})}{D(P_0||P_1)}}{\alpha(1-\pi)+\pi(1-\beta)} \\ &\leq \frac{C_2' + \log\left(\frac{1-\pi}{\pi\delta}\right)}{D(P_1||P_0)} + \frac{C_1' + \log\gamma_{\mathrm{L}}^{-1}}{\pi(1-\gamma_{\mathrm{L}})D(P_0||P_1)} \\ &\leq \frac{C_1}{\pi D(P_0||P_1)} + \frac{\log\frac{1}{\pi\delta}}{D(P_1||P_0)} + \frac{C_2}{D(P_1||P_0)}\end{aligned}$$

where the second inequality follows from dropping $\alpha(1-\pi)$ from the denominator, replacing β with the bound in (19), and dropping $(1-\alpha)(1-\pi)$ from the numerator of the second term. The third inequality follows from defining $C_2 = C_2'$, and

$$C_1 = \frac{C_1' + \log\gamma_{\mathrm{L}}^{-1}}{1-\gamma_{\mathrm{L}}}. \tag{24}$$

■

Example 1. Searching for a Gaussian with positive mean. Consider searching for a population following $P_1 \sim \mathcal{N}(\mu, 1)$ amongst a number of populations following $P_0 \sim \mathcal{N}(-\mu, 1)$ for some $\mu > 0$. The Kullback-Leibler divergence between the Gaussian distributions is $D(P_0||P_1) = 2\mu^2$. From [19], p.145, we have an explicit expression for C_1':

$$C_1'(\mu) = 2\mu\left(\mu + \frac{e^{-\mu^2/2}}{\int_{-\mu}^{\infty} e^{-t^2/2}dt}\right). \tag{25}$$

In order to make our bound on $\mathbb{E}[N]$ as tight as possible, we would like to minimize C_1 from (24) with respect to γ_{L}. Since the minimizer has no closed form expression, we use the sub-optimal value $\gamma_{\mathrm{L}} = 1/\mu$, for $\mu \geq 1$, and $\gamma_{\mathrm{L}} = \mu$ for $\mu < 1$. For this choice of γ_{L}, the constant $C_1 = C_1(\mu)$ in Theorem 7 and (24) is

$$C_1(\mu) = \begin{cases} \frac{C_1'(\mu)+\log(\mu)}{1-1/\mu} & \text{if } \mu \geq 1 \\ \frac{C_1'(\mu)+\log(1/\mu)}{1-\mu} & \text{if } \mu < 1. \end{cases}$$

Consider the following two limits. First, as $\mu \to \infty$

$$\lim_{\mu\to\infty} \frac{C_1(\mu)}{D(P_0||P_1)} = 1.$$

As a consequence (from Corollary 6)

$$\lim_{\mu\to\infty}\lim_{\pi\to 0} \pi\,\mathbb{E}[N] \leq \lim_{\mu\to\infty} \frac{C_1(\mu)}{D(P_0||P_1)} = 1.$$

Corollary 3 implies this bound is tight. As μ tends to infinity we approach the noise-free case, and the procedure is able to make perfect decisions with one sample per population. As expected, the required number of samples grows as $1/\pi$.

Second, as $\mu \to 0$,

$$\lim_{\mu\to 0} C_1(\mu) = 1$$

which implies (again from Corollary 6)

$$\lim_{\mu\to\infty}\lim_{\pi\to 0} \pi\, D(P_0||P_1)\,\mathbb{E}[N] \leq \lim_{\mu\to\infty} C_1(\mu) = 1.$$

Comparison to Corollary 4 shows the bound is tight. For small π, the S-SPRT requires $1/(\pi D(P_0||P_1))$ samples as the distributions become very similar, and no procedure can do better.

Fig. 1 plots the expected number of samples scaled by π as a function of μ. Specifically, the figure shows three plots. First, μ vs. $\pi\,\mathbb{E}[N]$ obtained from simulation of the S-SPRT procedure with $\pi = 10^{-3}$, $\gamma_{\mathrm{L}} = 1$ and $\gamma_{\mathrm{U}} = \frac{1-\pi}{\pi\delta}$ for $\delta = 10^{-2}$ is plotted. Second, the lower bound from Theorem 5 is shown. For small π, from (7), any reliable procedure has

$$\mathbb{E}[N] \gtrsim \frac{1}{\pi}\max\left(1, \frac{1}{D(P_0||P_1)}\right).$$

Lastly, the upper bound from Theorem 7 is plotted. From (17), for small values of π, the S-SPRT achieves

$$\mathbb{E}[N] \lesssim \frac{C_1}{\pi D(P_0||P_1)}.$$

where C_1 is calculated by minimizing (24) over values of $A \in (0,1)$ for each value of μ. C_1 is within a small factor of the lower bound for all values of μ.

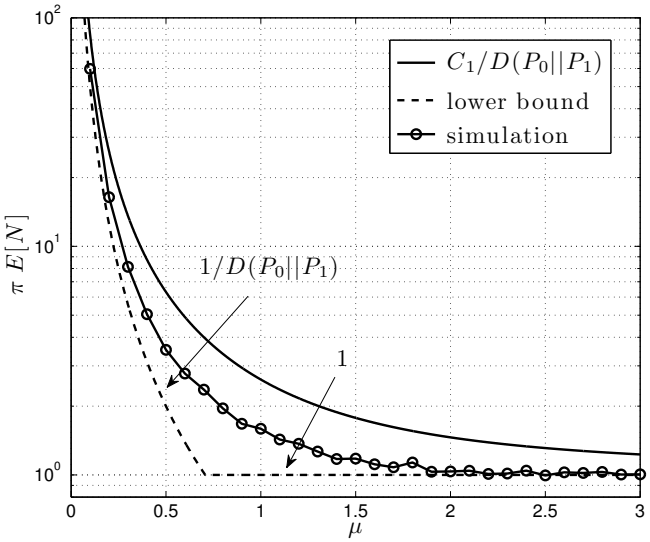

Fig. 1. Expected number of samples scaled by π as a function of the mean of the atypical population, μ, corresponding to example 1. Simulation of the S-SPRT is plotted with the upper bound from Corollary 6 and lower bound from Corollary 2. Simulation details: $\pi = 10^{-3}$, $P_e \leq 10^{-2}$, 10^3 trials for each value of μ.

Example 2. Searching for a biased coin. Consider the problem of searching for a coin with bias towards heads of $1/2 + b$ amongst coins with bias towards heads of $1/2 - b$, for $b \in [0, 1/2]$. This problem was studied recently in [9].

Corollary 8. Biased Coin. *The S-SPRT procedure with* $\gamma_{\mathrm{L}} = \frac{1-2b}{1+2b}$ *and* $\gamma_{\mathrm{U}} = \frac{1-\pi}{\pi\delta}$ *satisfies* $P_e \leq \frac{\delta}{1+\delta}$ *and*

$$\mathbb{E}[N] \leq \frac{1}{2b^2}\left(\frac{1}{\pi} + \log\left(\frac{1}{\pi\delta}\right) + 1\right).$$

Proof: The proof follows from evaluation of the constants in Theorem 7. The log-likelihood ratio corresponding to each sample (each coin flip) takes one of two values: if a coin reveals heads, $L^{(1)} = \log \frac{1+2b}{1-2b}$, and if a coin reveals tails, $L^{(1)} = \log \frac{1-2b}{1+2b}$. When each individual SPRT terminates, it can exceed the threshold by no more than this value, giving,

$$C_1'(b) = \log \frac{1+2b}{1-2b} \qquad C_2'(b) = \log \frac{1+2b}{1-2b}.$$

With $\gamma_{\mathrm{L}} = \frac{1-2b}{1+2b}$, we can directly calculate the constants in Theorem 7. From (24),

$$\frac{C_1(b)}{D(P_0||P_1)} = \frac{1+2b}{4b^2} \leq \frac{1}{2b^2}$$

as the Kullback-Leibler divergence is $D(P_0||P_1) = D(P_1||P_0) = 2b \log \frac{1+2b}{1-2b}$. Also note

$$\frac{1}{D(P_1||P_0)} \leq \frac{1}{2b^2}.$$

Lastly,

$$\frac{C_2(b)}{D(P_1||P_0)} = \frac{C_2'}{D(P_1||P_0)} = \frac{1}{2b} \leq \frac{1}{2b^2}.$$

Combining these with Theorem 7 completes the proof. ∎

Comparison of Corollary 8 to Theorem 2 of [9] shows the leading constant is a factor of 32 smaller in the bound presented here.

Moreover, closer inspection reveals that the constant $C_1(b)$ can be further tightened. Specifically, note that when an individual SPRT estimates $\hat{X}_i = 0$ it must hit the lower threshold exactly (since $\gamma_L = (1-2b)/(1+2b)$). If we choose only values of δ such that the upper threshold is an integer multiple of the likelihood ratio (i.e., set $\log \gamma_{\mathrm{U}} = k \log((1+2b)/(1-2b))$ for some integer k) the overshoot here is also zero. $C_1' = 0$ and $C_2' = 0$, which then give

$$\frac{C_1(b)}{D(P_0||P_1)} = \frac{1+2b}{8b^2}.$$

From Corollary 6,

$$\lim_{\pi \to 0} \pi\, \mathbb{E}[N] \leq \frac{1+2b}{8b^2}. \tag{26}$$

For small π, the number of samples required by any procedure to reliably identify an atypical population is

$$\mathbb{E}[N] \gtrsim \frac{1}{\pi} \left(\frac{1+2b}{8b^2} \right).$$

If $b = 1/2$ (each coin flip is deterministic), $C_1/D(P_0||P_1) = 1$, and the expected number of samples grows as $1/\pi$ as expected. The upper bound in Corollary 6 and lower bound in Corollary 1 converge.

Likewise, as the bias of the coin becomes small, $\lim_{b \to 0} C_1(b) = 1$, and the expected number of samples to reliably identify an atypical population grows as $1/(\pi D(P_0||P_1))$. Again the upper and lower bounds converge.

Note that the S-SPRT procedure for testing the coins is equivalent to a simple, intuitive procedure, which can be implemented as follows: beginning with coin i, and a scalar static $T = 0$, if heads appears, add 1 to the statistic. Likewise, if tails appears, subtract 1 from the test statistic. Continue to flip the coin until either *1)* T falls below 0, or *2)* T exceeds some upper threshold (which controls the error rate). If the statistic falls below 0, move to a new coin, and reset the count, i.e., set $T = 0$; conversely if the statistic exceeds the upper threshold, terminate the procedure. Note that any time the coin shows tails on the first flip, the procedure immediately moves to a new coin.

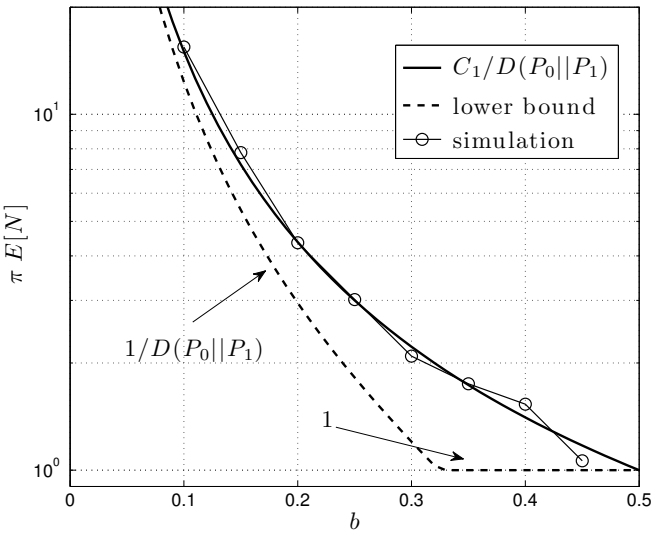

Fig. 2. Expected number of samples scaled by π as a function of the bias of the coin corresponding to Example 2. Upper and lower bounds from Corollaries 6 and 2. Simulation of the S-SPRT: $\pi = 10^{-3}$, $P_e \leq 10^{-2}$, 10^2 trials for each value of b.

Fig. 2 plots the expected number of samples scaled by π as a function of the bias of the atypical coins, b. The S-SPRT was simulated with the lower threshold set at $\gamma_{\mathrm{L}} = \log \frac{1-2b}{1+2b}$ for $\pi = 10^{-3}$ and $P_e \leq 10^{-2}$. The upper and lower bounds from Corollaries 6 and 1 are also plotted. The constant in the upper bound is given by the expression in (26).

Notice that the simulated procedure appears to achieve the upper bound. Closer inspection of the derivation of Theorem 7 with $C_1' = 0$ (as the *overshoot* in (22) is zero), shows the bound on the number of samples required by the S-SPRT is indeed tight for the search for the biased coin.

Remark 1. The S-SPRT procedure is fairly insensitive to our knowledge of the true prior probability π. On one hand, if we overestimate π by using a larger $\tilde{\pi}$ to specify the upper threshold $\gamma_{\mathrm{U}} = \frac{1-\tilde{\pi}}{\tilde{\pi}\delta}$, then according to (20) the probability of error P_e increases and is approximately $\frac{\tilde{\pi}}{\pi} \frac{\delta}{1+\delta}$, while the order of $\mathbb{E}[N]$ remains the same. On the other hand, if our $\tilde{\pi}$ underestimates π, then the probability of error P_e is reduced by a factor of $\tilde{\pi}/\pi$, and the order of $\mathbb{E}[N]$ also remains the same, provided $\log(1/\tilde{\pi}) \leq 1/\pi$, i.e., $\tilde{\pi}$ is not exponentially smaller than π. As a consequence, it is sensible to underestimate π,

rather than overestimate π as the latter would increase the probability of error.

Remark 2. Implementing a sequential probability ratio test on each population can be challenging for many practical problems. While the S-SPRT is optimal when both P_0 and P_1 are known and testing a single population amounts to a simple binary hypothesis test, scenarios often arise where some parameter of distribution P_1 is unknown. Since the SPRT is based on exact knowledge of P_1, it cannot be implemented in this case. Many alternatives to the SPRT have been proposed for *composite* hypothesis (see [20], [21], etc.). In the next section we propose an alternative that is near optimal and also very simple to implement.

V. Sequential Thresholding

Sequential thresholding, first proposed for sparse recovery problems in [7], can be applied to the search for an atypical population, and admits a number of appealing properties. It is particularly well suited for problems in which the atypical distributions are rare. While sequential thresholding requires more samples than the S-SPRT, it does not require full knowledge of the distributions, specifically P_1, as required by the S-SPRT (see Remarks 2 and 4). Moreover, the procedure admits a general error analysis, and perhaps most importantly is very simple to implement (a similar procedure is used in the SETI project [5], [6]). Somewhat surprisingly, the procedure can substantially outperform non-adaptive procedures as π becomes small. Roughly speaking, for small values of π, the procedure reliably recovers an atypical population with

$$\mathbb{E}[N] \lesssim \frac{\log\log \pi^{-1}}{\pi D(P_0||P_1)}.$$

Algorithm 3 Sequential Thresholding

input: integer $k_{\max}$, integer m, threshold γ
initialize: $i = 1$, $k = 1$
while $k < k_{\max} + 1$ **do**
  measure: $(Y_{i,(k-1)m+1}, ..., Y_{i,km})$
  if $T\left(Y_{i,(k-1)m+1}, ..., Y_{i,km}\right) \leq \gamma$ **then**
    $i = i + 1$
    $k = 1$
  else
    $k = k + 1$
  end if
end while
output: $\hat{X}_i = 1$

Sequential thresholding requires three inputs: *1)* $k_{\max}$, an integer representing the maximum number of rounds for any particular index, *2)* m, an integer representing the number of samples per round and *2)* γ, a threshold. Let $T(Y_{i,1}, ..., Y_{i,m})$ represent a sufficient statistic that does not depend on the parameters of P_1 or P_0 (for example, in the Gaussian case, $T(Y_{i,1}, ..., Y_{i,m}) = \sum_{j=1}^{m} Y_{i,j}$).

The procedure searches for an atypical population as follows. Starting on population i, the procedure takes m samples. If the sufficient statistic comprised of those m samples is greater than the threshold, i.e. $T(Y_{i,1}, ..., Y_{i,m}) > \gamma$, the procedures takes an additional *block* of m samples of index i and forms $T(Y_{i,2m+1}, ..., Y_{i,2m})$ (which is only a function of the *second* block of m samples). If $T(Y_{i,2m+1}, ..., Y_{i,2m}) > \gamma$, a third block of samples is taken. The procedure continues in this manner, re-testing the statistic *up to* a maximum of $k_{\max}$ times. If the statistic is below the threshold, i.e. $T < \gamma$, after any sample, the procedure immediately moves to the next population, setting $i = i + 1$, and resetting k. Should any population survive all $k_{\max}$ rounds, the procedure estimates $\hat{X}_i = 1$, and terminates. The procedure is described in Algorithm 3.

Control of the probability of error depends on the threshold γ, the number of rounds $k_{\max}$, and the number of samples per round, m. Define the probability that the test statistic is below the threshold given the current index follows P_0 as

$$\rho := \mathbb{P}_0\left(T > \gamma\right)$$

where $\rho \in (0, 1)$. Note that ρ is fixed and not a function of π.

Intuitively, the procedure can control the probability of error as follows. First, α can be made small by simply increasing $k_{\max}$, as, by the independence of the blocks of samples, $\alpha = \rho^{k_{\max}}$. Of course, as $k_{\max}$ is increased, β also increases. In order to control β, m is increased. As we show in the following theorem, to control β, it is sufficient to have m grow as $\log\log \pi^{-1}$. This $\log\log \pi^{-1}$ can be interpreted as the penalty the sub-optimal procedure pays for increased robustness. The following theorem quantifies the number of samples required to recover an index following P_1.

Theorem 9. Sequential Thresholding. *For any* $\rho \in (0, 1)$, $\delta \in (0, 1)$, *and* $\epsilon > 0$ *set* $k_{\max} = \left\lceil \log_{1/\rho}\left(\frac{1-\pi}{\pi\delta}\right)\right\rceil$ *and* $m = \left\lceil \frac{(1+\epsilon)\log k_{\max}}{D(P_0||P_1)}\right\rceil$. *Sequential thresholding then satisfies*

$$\lim_{\pi\to 0} P_e \leq \frac{\delta}{1+\delta}$$

and

$$\lim_{\pi\to 0} \frac{\pi}{\log\log_{1/\rho} \pi^{-1}}\, \mathbb{E}[N] \leq \frac{1+\epsilon}{D(P_0||P_1)(1-\rho)}.$$

Proof: Employing sequential thresholding, the false positive event depends on the number of rounds as $\alpha = \rho^{k_{\max}}$. With $k_{\max}$ as specified, we have $\alpha \leq \pi\delta/(1-\pi)$. The probability the procedure returns a population corresponding to P_0 then follows from (4) as

$$P_e \leq \frac{\delta}{\delta + 1 - \beta}. \tag{27}$$

Next, we show β tends to zero as π becomes small. The Chernoff-Stein Lemma [22] states that since ρ is fixed,

$$\lim_{m\to\infty} \frac{\log \mathbb{P}_1\left(T \leq \gamma\right)}{m} = -D(P_0||P_1),$$

which implies

$$\lim_{\pi\to 0}\frac{\log \mathbb{P}_1(T\le\gamma)}{\log k_{\max}} = -(1+\epsilon)$$

since $m = \left\lceil \frac{(1+\epsilon)\log k_{\max}}{D(P_0||P_1)} \right\rceil$ and $\lim_{\pi\to 0} m = \infty$. By definition,

$$\beta = \mathbb{P}_1\left(\bigcup_{k=1}^{k_{\max}} T\le\gamma\right) \le k_{\max}\mathbb{P}_1(T\le\gamma).$$

where the inequality holds by the union bound. Therefore, we obtain

$$\begin{aligned}\lim_{\pi\to 0}\beta &\le \lim_{\pi\to 0}\exp\left(\log(k_{\max})\left(1+\frac{\log \mathbb{P}_1(T\le\gamma)}{\log k_{\max}}\right)\right)\\ &= \exp\left(-\epsilon\lim_{\pi\to 0}\log k_{\max}\right)\\ &= 0.\end{aligned}$$

Combined with (27), we have

$$\lim_{\pi\to 0} P_e \le \frac{\delta}{1+\delta}.$$

The expected number of samples required for any index following P_0 is given as

$$E_0 = \sum_{k=1}^{k_{\max}} m\,\rho^{k-1} \le \frac{m}{1-\rho}.$$

On the other hand, the expected number of samples given the index follows P_1 is less than m times the maximum number of rounds:

$$E_1 \le m k_{\max}.$$

From (3) we have

$$\begin{aligned}\mathbb{E}[N] &\le \frac{\pi m k_{\max} + (1-\pi)\frac{m}{1-\rho}}{\alpha(1-\pi)+\pi(1-\beta)}\\ &\le \frac{m k_{\max}}{(1-\beta)} + \frac{m(1-\pi)}{\pi(1-\rho)(1-\beta)}.\end{aligned}$$

With $k_{\max}$ and m as specified,

$$\lim_{\pi\to 0}\frac{\pi}{\log\log_{1/\rho}\pi^{-1}}\,\frac{m k_{\max}}{1-\beta} = 0$$

and

$$\begin{aligned}\lim_{\pi\to 0}\frac{\pi}{\log\log_{1/\rho}\pi^{-1}}\frac{m(1-\pi)}{\pi(1-\rho)(1-\beta)}\\ = \frac{1+\epsilon}{D(P_0||P_1)(1-\rho)}.\end{aligned}$$

implying

$$\lim_{\pi\to 0}\frac{\pi}{\log\log_{1/\rho}\pi^{-1}}\mathbb{E}[N] \le \frac{1+\epsilon}{D(P_0||P_1)(1-\rho)}.$$

■

Remark 3. Similar to the behavior of the SPRT discussed in Remark 1, sequential thresholding is also fairly insensitive to our prior knowledge of π, especially when we underestimate π. More specifically, overestimating π increases the probability of error almost proportionally and has nearly no affect on $\mathbb{E}[N]$, while underestimating π decreases the probability of error and the order of $\mathbb{E}[N]$ is the same as long as $\log(1/\tilde{\pi}) \le 1/\pi$.

Remark 4. For many distributions in the exponential family, the log-likelihood ratio, L_i, defined in (21) is a monotonic function of a test statistic T that does not depend on parameters of P_1. As a consequence of the sufficiency of T, the threshold γ depends only on P_0, making sequential thresholding suitable when knowledge about P_1 is not available.

Perhaps most notably, in contrast to the SPRT based procedure, sequential thresholding does not aggregate statistics. Roughly speaking, this results in increased robustness to modeling errors in P_1 at the cost of a sub-optimal procedure. Analysis of sequential thresholding in related sparse recovery problems can be found in [7], [11].

VI. Limitations of Non-Adaptive Procedures

For our purposes a non-adaptive procedure tests each individual population with a pre-determined number of samples, denoted N_0. In this case, the conditional number of samples for each individual test is simply $E_0 = E_1 = N_0$ giving

$$\mathbb{E}[N] = \frac{N_0}{\alpha(1-\pi)+\pi(1-\beta)}. \tag{28}$$

To compare the sampling requirements of non-adaptive procedures to adaptive procedures, we present a necessary condition for reliable recovery. The theorem implies that non-adaptive procedures require a factor of $\log\pi^{-1}$ more samples than the best adaptive procedures.

Theorem 10. Non-adaptive procedures. *Any non-adaptive procedure that satisfies*

$$P_e \le \frac{\delta}{1+\delta}$$

also has

$$\mathbb{E}[N] \ge \frac{\log\left(\frac{1}{2\delta\pi}\right)-1}{\pi(1+\delta)D(P_1||P_0)}.$$

for $\delta \le 1/2$.

Proof: Assume that $P_e \le \frac{\delta}{1+\delta}$ and from (4) we have

$$\frac{\alpha(1-\pi)}{\pi(1-\beta)} \le \delta. \tag{29}$$

From (28),

$$\mathbb{E}[N] \ge \frac{N_0}{\pi(1+\delta)(1-\beta)}.$$

Next, for any binary hypothesis test with false negative α and false positive β, the following identity holds:

$$N_0 \ge \frac{\beta\log\left(\frac{\beta}{1-\alpha}\right)+(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right)}{D(P_1||P_0)}. \tag{30}$$

To see (30), recall that for non-adaptive procedures, $N_0 = E_0 = E_1$, and thus both bounds in (9) and (10) apply. This gives

$$\begin{aligned}\mathbb{E}[N] &\geq \frac{\beta \log\left(\frac{\beta}{1-\alpha}\right)}{\pi(1+\delta)(1-\beta)D(P_1||P_0)} + \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\pi(1+\delta)D(P_1||P_0)} \\ &\geq \frac{\log\left(\frac{1-\pi}{\delta\pi}\right) - 1}{\pi(1+\delta)D(P_1||P_0)} \\ &\geq \frac{\log\left(\frac{1}{2\delta\pi}\right) - 1}{\pi(1+\delta)D(P_1||P_0)}\end{aligned}$$

where the second inequality follows from (11) and (29), and the last inequality as $\pi \leq 1/2$.

■

Remark 1. The lower bound presented in Theorem 10 implies that non-adaptive procedures require at best a multiplicative factor of $\log \pi^{-1}$ more samples than adaptive procedures (as adaptive procedures are able to come within a small constant of the lower bound in Theorem 5). For problems with even modestly small values of π, this results in non-adaptive sampling requirements many times larger than those required by adaptive sampling procedures.

VII. Conclusion

This paper explored the problem of finding an atypical population amongst a number of typical populations, a problem arising in many aspects of science and engineering.

More specifically, this paper quantified the number of samples required to recover an atypical population with high probability. We paid particular attention to problems in which the atypical populations themselves become increasingly rare. After establishing a lower bound based on the Kullback Leibler divergence between the underlying distributions, the number of samples required by the optimal S-SPRT procedure was studied; the number of samples is within a constant factor of the lower bound, which can be explicitly derived in a number of cases. Two common examples, where the distributions are Gaussian and Bernoulli, were studied.

Sequential thresholding, a more robust procedure that can be implemented with less prior knowledge about the distributions, was presented and analyzed in the context of the quickest search problem. Sequential thresholding requires a multiplicative factor more samples, doubly logarithmic in the prior, than the S-SPRT procedure. Both sequential thresholding and the SPRT procedure were shown to be fairly robust to modeling errors in the prior probability. Lastly, for comparison, a lower bound for non-adaptive procedures was presented.

References

[1] M. Malloy, G. Tang, and R. Nowak, "Quickest search for a rare distribution," in *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, march 2012.

[2] A. Tajer, R. Castro, and X. Wang, "Adaptive sensing of congested spectrum bands," *Arxiv preprint arXiv:1206.1438*, 2012.

[3] W. Zhang, A. Sadek, C. Shen, and S. Shellhammer, "Adaptive spectrum sensing," in *Information Theory and Applications Workshop (ITA), 2010*. IEEE, 2010, pp. 1–7.

[4] L. Lai, H. Poor, Y. Xin, and G. Georgiadis, "Quickest sequential opportunity search in multichannel systems," in *Proc. Int. Workshop Applied Probability*.

[5] J. H. Wolfe, J. Billingham, R. E. Edelson, R. B. Crow, S. Gulkis, and E. T. Olsen, "SETI - the search for extraterrestrial intelligence - plans and rationale," *Life in the Universe. Proceedings of the Conference on Life in the Universe, NASA Ames Research Center*, 1981.

[6] D. Overbye, "Search for aliens is on again, but next quest is finding money," *The New York Times*, 2012.

[7] M. Malloy and R. Nowak, "Sequential analysis in high-dimensional multiple testing and sparse recovery," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, 31 2011-aug. 5 2011, pp. 2661 –2665.

[8] L. Lai, H. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *Information Theory, IEEE Transactions on*, vol. 57, no. 8, pp. 5375 –5386, aug. 2011.

[9] K. Chandrasekaran and R. Karp, "Finding the most biased coin with fewest flips," *ArXiv e-prints*, Feb. 2012.

[10] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," *http://arxiv.org/abs/1001.5311*, 2010.

[11] M. Malloy and R. Nowak, "On the limits of sequential testing in high dimensions," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, nov. 2011, pp. 1245 –1249.

[12] M. Marcus and P. Swerling, "Sequential detection in radar with multiple resolution elements," *Information Theory, IRE Transactions on*, vol. 8, no. 3, pp. 237 –245, april 1962.

[13] E. Posner and J. Rumsey, H., "Continuous sequential decision in the presence of a finite number of hypotheses," *Information Theory, IEEE Transactions on*, vol. 12, no. 2, pp. 248 – 255, apr 1966.

[14] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Algorithmic Learning Theory*. Springer, 2009, pp. 23–37.

[15] S. Mannor and J. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *The Journal of Machine Learning Research*, vol. 5, pp. 623–648, 2004.

[16] D. Siegmund, *Sequential Analysis*. New York, NY, USA: Springer-Verlag, 2010.

[17] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.

[18] M. N. Ghosh, "Bounds for the expected sample size in a sequential probability ratio test," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 22, no. 2, pp. pp. 360–367, 1960. [Online]. Available: http://www.jstor.org/stable/2984106

[19] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. pp. 117–186, 1945. [Online]. Available: http://www.jstor.org/stable/2235829

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

[21] G. Fellouris and A. Tartakovsky, "Almost minimax sequential tests of composite hypotheses," *Arxiv preprint arXiv:1204.5291*, 2012.

[22] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2005.