Quickest Search for a Rare Distribution

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*—We consider the problem of finding one sequence of independent random variables following a rare atypical distribution, P_1, amongst a large number of sequences following some null distribution, P_0. We quantify the number of samples needed to correctly identify one atypical sequence as the atypical sequences become increasingly rare. We show that the known optimal procedure, which consists of a series of sequential probability ratio tests, succeeds with high probability provided the number of samples grows at a rate equal to a constant times $\pi^{-1}D(P_1||P_0)^{-1}$, where π is the prior probability of a sequence being atypical, and $D(P_1||P_0)$ is the Kullback-Leibler divergence. Using techniques from sequential analysis, we show that if the number of samples grow at a rate equal to $\pi^{-1}D(P_1||P_0)^{-1}$, any procedure fails. This is then compared to sequential thresholding [1], a simple procedure which can be implemented without exact knowledge of distribution P_1. We also show that the SPRT and sequential thresholding are fairly robust to our knowledge of π. Lastly, a lower bound for non-sequential procedures is derived for comparison.**

***Index Terms*—Rare events, SPRT, CUSUM procedure, sparse recovery, sequential analysis, sequential thresholding, spectrum sensing.**

I. Introduction

This paper studies the problem of identifying a rare, atypical distribution amongst a large number of typical distributions. Consider a number of random sequences, indexed by $i = 1, 2, ...$ The samples of each sequence are independent and identically distributed according to one of two distributions:

$$Y_i \sim \begin{cases} P_0 & \text{w.p.} \quad (1-\pi) \\ P_1 & \text{w.p.} \quad \pi. \end{cases} \qquad i = 1, 2, ... \qquad (1)$$

The prior probability that any sequence is atypical is denoted π, and, as we are interested in the rare event regime, π is small. The goal of the problem is to find one such atypical sequence using as few samples as possible. This paper presents results quantifying the number of samples required as π becomes vanishingly small.

As the occurrence of sequences following P_1 becomes increasingly rare (i.e. as $\pi \to 0$), the number of samples required to find one such sequence must, of course, increase. If P_0 and P_1 are extremely different (e.g., non-overlapping supports), then a testing procedure could simply proceed by taking one sample for each i until an atypical sequence was found. The procedure would identify an atypical sequence using on the order of π^{-1} samples. More generally, when the two distributions are more difficult to distinguish, we must take multiple samples of some sequences. In this case, a cumulative sum (CUSUM) test, which consists of a series of sequential probability ratio tests (SPRT), is optimal [2].

Finding an atypical distribution from observations 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 amongst 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.

Another captivating example is that of 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"). Roughly speaking, SETI employs a variable sample size search procedure that repeatedly tests energy levels against a threshold up to five times [3], [4]. If any of the measurements are below the threshold, the procedure immediately passes to the next frequency/star. This procedure is closely related to that of *Sequential Thresholding* [1]. Sequential Thresholding results in substantial gains over fixed sample size procedures and, unlike the SPRT, it can be computed without perfect knowledge of P_1.

Prior work most closely related to the problem studied here is that by Lai, Poor, Xin, and Georgiadis [2], in which the authors examine the same problem, but keep π fixed. The authors show that a CUSUM test is optimal. In this case, the CUSUM procedure is equivalent to a series of SPRTs, one applied to each sequence. Note the implementation of CUSUM requires knowledge of P_1, which is often not available in practice. This problem is also studied in [5] in which the distributions are restricted to Bernoulli random variables. The problem is closely related to sparse signal support recovery from point-wise observations [1], [6], [7]. The sparse signal recovery problem differs in that the objective is to recover all (or most) sequences following P_1, as opposed to finding a single sequence and terminating the procedure.

The main contributions of this paper are to quantify the minimum search times required by several different tests in the small π regime. The results presented in this paper are in terms of the asymptotic rate at which the expected number of measurements must grow as π becomes small, in order to guarantee the procedure finds a sequence following distribution P_1. We first show that the SPRT-based procedure can succeed with probability tending to one provided the number of measurements grows at a rate equal to a constant times $\pi^{-1}D(P_1||P_0)^{-1}$. We give an explicit expression for this constant in the Gaussian case. Next we explore two lower bounds. When $D(P_1||P_0) < 1$, we show that any sequential procedure fails with probability bounded away from zero provided the expected number of samples is less than $\pi^{-1}D(P_1||P_0)^{-1}$; when $D(P_1||P_0) \geq 1$ any procedure fails if the number of samples grows slower than π^{-1}. We compare these results to sequential thresholding, a simple sequential procedure that, unlike the SPRT, does not require perfect knowledge of P_1. We also show that both the SPRT and sequential thresholding are fairly insensitive to our knowledge of the prior probability π. Lastly, for comparison, we derive and discuss the performance limitations of fixed sample size procedures. Fixed sample size procedures take the same number of samples for each sequence, and require a factor of $\log \pi^{-1}$ more samples.

II. Problem Setup

Consider an infinite number of sequences indexed by $i = 1, 2, ...$ For each index i, samples 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 with joint support on $\mathcal{Y}$, j indexes multiple i.i.d. samples of a particular sequence, and x is a binary label. The goal is to find a sequence i such that $x_i = 1$ as quickly and reliably as possible. The prior probability of a particular index i following P_1 or P_0 is denoted

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

Without loss of generality, a testing procedure starts at index $i = 1$ and takes one sample. The procedure then decides to either *1)* take an additional sample of index $i = 1$, or *2)* estimate sequence $i = 1$ as following distribution P_0 (deciding $\hat{x}_1 = 0$) and move to index 2, or *3)*, terminate, declaring sequence $i = 1$ 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 [2], the procedure does not revisit sequences (a well justified assumption as each sequence 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 P_1, which we denote P_e,

$$P_e := \mathbb{P}(I \in \{i : x_i = 0\})$$

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

Imagine that the procedure is currently sampling index i. Agnostic to the particular sampling procedure used, 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 index 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 expected number of measurements required by the procedure can be expressed as follows. Let N_i be the number of samples taken on index 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 measurements as

$$\begin{aligned} \mathbb{E}[N] &= \mathbb{E}[N_1] \quad + \\ &\quad \mathbb{E}[N_2 + N_3 + ... \ |\hat{x}_1 = 0] \left((1-\pi)(1-\alpha) + \pi\beta\right). \end{aligned} \tag{2}$$

The expected number of samples used from the second index onwards, given that we arrive at the second index, 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 (2) gives

$$\mathbb{E}[N] = \frac{\mathbb{E}[N_1]}{\alpha(1-\pi) + \pi(1-\beta)}. \tag{3}$$

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

$$E_1 = \mathbb{E}[N_1 | x_1 = 1] \qquad E_0 = \mathbb{E}[N_1 | x_1 = 0].$$

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

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

We are interested in the rate at which $\mathbb{E}[N]$ must grow as π becomes small, i.e, as the occurrence of sequences following distribution P_1 becomes increasingly rare, to ensure that P_e goes to zero. In general, α, β, P_e and $\mathbb{E}[N]$ are functions of π, though we suppress this dependence for notational convenience. We can write (4) as

$$P_e = \frac{1}{1 + \frac{\pi(1-\beta)}{\alpha(1-\pi)}}.$$

From this expression we see that if

$$\lim_{\pi \to 0^+} \frac{\alpha(1-\pi)}{\pi(1-\beta)} = 0$$

then P_e goes to zero as π becomes small. Likewise, if

$$\lim_{\pi \to 0^+} \frac{\alpha(1-\pi)}{\pi(1-\beta)} \geq \delta \tag{5}$$

for some $\delta > 0$, then

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

and, in the limit, P_e is greater than or equal to some positive constant.

III. SPRT Procedure

The SPRT, optimal for simple binary hypothesis tests, can be applied to the problem studied here by implementing a series of SPRTs on the individual sequences. This is equivalent in form to the CUSUM test proposed in [2], which is traditionally applied to change point detection problems. Imagine the procedure has currently taken j samples of sequence i. The procedure continues to sample sequence i provided

$$A \leq \prod_{k=1}^{j} \frac{P_1(Y_{i,k})}{P_0(Y_{i,k})} \leq B. \tag{6}$$

In words, the procedure continues to sample sequence i provided the likelihood ratio comprised of samples of that sequence is between two scalar thresholds. The procedure stops sampling a particular sequence after N_i samples, which is a random integer representing the smallest number of samples such that (6) no longer holds:

$$N_i := \min \left\{ j : \prod_{k=1}^{j} \frac{P_1(Y_{i,k})}{P_0(Y_{i,k})} < A \bigcup \prod_{k=1}^{j} \frac{P_1(Y_{i,k})}{P_0(Y_{i,k})} > B \right\}.$$

When the likelihood ratio exceeds B, then $\hat{x}_i = 1$, and the procedure terminates returning $I = i$. Conversely, if the likelihood ratio falls below A, then $\hat{x}_i = 0$, and the procedure moves to index $i+1$.

The SPRT procedure studied in [2] fixes the lower threshold in each individual SPRT at $A = 1$, which has a very intuitive interpretation. Since there are an infinite number of sequences, anytime a sample suggests that a particular sequence doesn't follow P_1, moving to another sequences is best. While this approach is optimal [2], we use a strictly smaller threshold, as it results in a simpler derivation of our bound. We continue with the first theorem of the paper.

Theorem 1. *Performance of SPRT procedure. Consider a series of SPRTs used to test for a sequence following P_1 with upper threshold $B = \frac{1-\pi}{\pi\delta}$ for $\delta > 0$, and fixed lower threshold $A \in (0,1)$. The procedure then satisfies $P_e \leq \frac{\delta}{1+\delta}$ and has*

$$\mathbb{E}[N] \leq \frac{C}{\pi D(P_0||P_1)}$$

for some sufficiently small π, and some constant C independent of π.

Proof: The proof is based on well known techniques used for analysis of the SPRT. From [8], the false positive and false negative events are related to the thresholds as:

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

$$\beta \leq A(1-\alpha) \leq A \tag{8}$$

It then follows from (4) that the probability the procedure terminates in error, returning a sequence following P_0, is

$$P_e = \frac{1}{1+\frac{\pi(1-\beta)}{\alpha(1-\pi)}} \leq \frac{1}{1+\frac{\pi}{B^{-1}(1-\pi)}} = \frac{\delta}{1+\delta}. \tag{9}$$

To show the second part of the theorem, we write the expected number of measurements as

$$\mathbb{E}[N] = \frac{\pi E_1 + (1-\pi)E_0}{\alpha(1-\pi)+\pi(1-\beta)}. \tag{10}$$

Define the log-likelihood ratio as

$$L_i^{(j)} = \sum_{k=1}^{j} \log \frac{P_1(Y_{i,k})}{P_0(Y_{i,k})}. \tag{11}$$

By Wald's identity [8],

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

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 [8]). In our case, we assume the value of the likelihood ratio when the procedure terminates can be bound by a constant. Specifically,

$$\mathbb{E}_0\left[L_i^{(N_i)}\middle|\hat{x}=0\right] \geq \log A - C_1$$

and

$$\mathbb{E}_0\left[L_i^{(N_i)}\middle|\hat{x}=1\right] \leq \log B + C_2$$

for some constants, $C_1, C_2 > 0$, independent of A and B. In practice, this is a minor assumption. C_1 and C_2 can be explicitly calculated for a variety of problems (see [9], p.145, and [10]). We have

$$\begin{aligned} E_0 &\leq \frac{(1-\alpha)\left(C_1 + \log A^{-1}\right) + \alpha\left(-C_2 + \log B^{-1}\right)}{D(P_0||P_1)} \\ &\leq \frac{C_1 + \log A^{-1}}{D(P_0||P_1)} \end{aligned} \tag{12}$$

where the second inequality follows as $B > A$. Likewise,

$$\begin{aligned} E_1 &\leq \frac{(1-\beta)(C_3 + \log B) + \beta(-C_4 + \log A)}{D(P_1||P_0)} \\ &\leq \frac{C_3 + \log B}{D(P_1||P_0)} \end{aligned} \tag{13}$$

for some constants $C_3, C_4 > 0$. Combining these with (10) gives

$$\begin{aligned}\mathbb{E}[N] &\leq \frac{\pi\frac{C_3+\log B}{D(P_1||P_0)} + (1-\pi)\frac{C_1+\log A^{-1}}{D(P_0||P_1)}}{\alpha(1-\pi)+\pi(1-\beta)} \qquad (14)\\ &\leq \frac{C_3+\log\left(\frac{1-\pi}{\pi\delta}\right)}{(1-A)D(P_1||P_0)} + \frac{(1-\pi)(C_1+\log A^{-1})}{\pi(1-A)D(P_0||P_1)}\end{aligned}$$

where the inequality follows from dropping $\alpha(1-\pi)$ from the denominator and replacing β with the bound in (7). The second term above dominates as $\pi \to 0$, and we have

$$\mathbb{E}[N] \leq (1+\epsilon)\frac{C}{\pi D(P_0||P_1)} \qquad (15)$$

for any $\epsilon > 0$ and π sufficiently small. As $1+\epsilon$ can be absorbed into the constant, this directly gives the theorem. Note that the constant C is an explicit function of the lower threshold and the constant C_1. Specifically,

$$C = \frac{C_1 + \log A^{-1}}{1-A}(1+\epsilon) \qquad (16)$$

where, again, C_1 is the overshoot of the log-likelihood ratio when the SPRT terminates. ∎

Remark 1. The 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 $B = \frac{1-\tilde{\pi}}{\tilde{\pi}\delta}$ in the SPRT procedure, then according to (9) 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 $\frac{\tilde{\pi}}{\pi}$, 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 better to underestimate π, rather than overestimate π as the latter would increase the probability of error.

Example 1. Testing Gaussians. Consider searching for one sequence following $P_1 \sim \mathcal{N}(0,\mu)$ amongst a large number of sequences following $P_0 \sim \mathcal{N}(0,-\mu)$. From [9], p.145, we have the following explicit expression for C_1

$$C_1(\mu) = 2\mu\left(\mu + \frac{e^{-\mu^2/2}}{\int_{-\mu}^{\infty} e^{-t^2/2}dt}\right).$$

Note that $\lim_{\mu\to\infty} C_1(\mu)/(2\mu^2) = 1$. In order to make $\mathbb{E}[N]$ as small as possible, ideally we would like to minimize C given in (28) with respect to A. Since the minimizer has no closed form expression, we use the sub-optimal value $A = 1/C_1(\mu)$. For this choice of A, the constant $C = C(\mu)$ in Theorem 1 is

$$C(\mu) = \frac{C_1(\mu) + \log(C_1(\mu))}{1 - 1/C_1(\mu)},$$

which has limit $\lim_{\mu\to\infty} C(\mu)/(2\mu^2) = 1$. A closer inspection shows that the minimal $C(\mu)$ optimized with respect to A also has this limit. For our example, the KL divergence between the Gaussian distributions is $D(P_0||P_1) = 2\mu^2$. As a consequence,

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

As μ tends to infinity we approach the noise-free case, and the procedure is able to make perfect decisions with one sample per sequence. As expected, the required number of samples tends to $1/\pi$.

IV. A Lower bound for any procedure

In this section we present conditions under which any sequential procedure fails to return a sequence following distribution P_1 with probability bounded away from zero.

Theorem 2. *Any (sequential) procedure with*

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

also has

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

Proof: We prove the theorem by contradiction. Assume that $\lim_{\pi\to 0^+} P_e < \frac{\delta}{1+\delta}$ and from (5) we have

$$\lim_{\pi\to 0^+} \frac{\alpha(1-\pi)}{\pi(1-\beta)} < \delta. \qquad (17)$$

This implies

$$\alpha(1-\pi) < \delta\pi(1-\beta) \qquad (18)$$

for all sufficiently small π. From (3),

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

From standard sequential analysis techniques [8] we have the following identity relating the expected number of measurements to α and β, which holds for any binary hypothesis testing procedure:

$$E_0 \geq \frac{\alpha\log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)}{D(P_0||P_1)}. \qquad (19)$$

Differentiating (19) with respect to α shows that the expression is monotonically decreasing in α over the set of α satisfying $\frac{\alpha}{1-\beta} < \frac{1-\alpha}{\beta}$. From (18), we are restricted to $\frac{\alpha}{1-\beta} < \frac{\delta\pi}{1-\pi}$ and thus, if $\frac{\delta\pi}{1-\pi} < \frac{1-\alpha}{\beta}$, then (19) is monotonically decreasing in α. We can replace α in (19) with $\alpha < \frac{\delta\pi(1-\beta)}{1-\pi}$. This gives

$$\begin{aligned}\mathbb{E}[N] &> \frac{(1-\pi)E_0}{\pi(1+\delta)(1-\beta)}\\ &> \frac{(1-\pi)\left(\alpha\log\left(\frac{\alpha}{1-\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)\right)}{\pi D(P_0||P_1)(1+\delta)(1-\beta)}\\ &> \frac{\delta\log\left(\frac{\delta\pi}{1-\pi}\right)}{(1+\delta)D(P_0||P_1)} +\\ &\quad \frac{(1-\pi)\left(1-\frac{\delta\pi(1-\beta)}{1-\pi}\right)\log\left(\frac{1}{\beta}-\frac{\delta\pi(1-\beta)}{\beta(1-\pi)}\right)}{\pi D(P_0||P_1)(1+\delta)(1-\beta)}\end{aligned}$$

Dropping all lower order terms as $\pi \to 0$, we have

$$\begin{aligned}\mathbb{E}[N] &> \frac{\log \frac{1}{\beta}}{(1+\delta)\pi D(P_0||P_1)(1-\beta)} \\ &> \frac{1}{(1+\delta)\pi D(P_0||P_1)} \end{aligned} \tag{20}$$

since $\frac{\log(1/\beta)}{1-\beta} \geq 1$.

Returning to the expression in (19), we also have the trivial inequality that $E_0 \geq 1$, giving an additional bound on the expected number of samples:

$$E[N] > \frac{(1-\pi)E_0}{(1+\delta)\pi(1-\beta)} \geq \frac{(1-\pi)}{(1+\delta)\pi}.$$

Thus, as π becomes sufficiently small, together with (20), the following holds

$$E[N] > \min\left(\frac{1}{(1+\delta)\pi}, \frac{1}{(1+\delta)\pi D(P_0||P_1)}\right)$$

and we have the theorem. ■

V. Sequential Thresholding

Implementing a sequential probability ratio test on each sequence can be challenging for a number of practical reasons. While the SPRT is optimal when both P_0 and P_1 are known and testing a single sequence amounts to a simple binary hypothesis test, scenarios often arise where some parameter of distribution P_1 is unknown. When some parameter of P_1 is unknown, the likelihood ratio cannot be formed, and sufficient statistics for the likelihood ratio result in adjustments to the thresholds based on the unknown parameters of distribution P_1. With incorrect thresholds, the SPRT is no longer optimal, and other issues arise (such as the *open continuation region* [8]). Another issue arising is robustness to outliers. When the model for distribution P_0 doesn't correctly account for large outliers, this can result in a large number of false positive events.

Sequential thresholding, first proposed for sparse recovery problems in [1], can be applied to the search for an atypical sequence, and admits a number of appealing properties. Specifically, it does not require full knowledge of the alternative distribution, admits a very general error analysis, and perhaps most importantly, is extremely simple to implement.

Algorithm 1

input: K steps, threshold γ
initialize: $i = 1$, $k = 1$
while $k < K + 1$ **do**
  measure: $\{Y_{i,j}\}_{j=1}^{m}$
  if $T_i^{(m)}(Y_{i,1}, ..., Y_{i,m}) \leq \gamma$ **then**
    $i = i + 1$, $k = 1$
  else
    $k = k + 1$
  end if
end while
output: $\hat{x}_i = 1$

Sequential thresholding requires two inputs: *1)* K, the number of passes, and *2)* γ, a threshold. Let $T_i^{(m)}$ be a sufficient statistic that does not depend on the parameters of P_1 or P_0. The probability a component following the null is eliminated on any given pass is related to the threshold as

$$\mathbb{P}_0\left(T_i^{(m)} \leq \gamma\right) = \frac{1}{2}.$$

The procedure operates by starting on sequence i, taking m samples. If $T_i^{(m)} > \gamma$, the procedures takes m more samples of i, forming and re-testing the likelihood ratio (using only these m samples). The procedure continues in this manner making *up to* a total of K passes. If $T_i^{(m)} < \gamma$ on any pass, $k = 1, ..., K$, the procedure immediately moves to the next index, setting $i = i + 1$, and resetting k. Should any index survive all K passes, the procedure estimates $\hat{x}_i = 1$, and terminates. The procedure is detailed above in Algorithm 1.

For many distributions in the exponential family, the log-likelihood ratio, $L_i^{(m)}$, defined in (11) is a monotonic function of $T_i^{(m)}$ under P_0. As a consequence of the sufficiency of $T_i^{(m)}$, the threshold γ depends only on P_0, making sequential thresholding more suitable when knowledge about P_1 is not available. Further details can be found in [1], [7].

Theorem 3. *Performance of Sequential Thresholding. Sequential Thresholding with* $K = \log_2\left(\frac{1-\pi}{\pi^{1+\delta}}\right)$ *succeeds in recovering a sequence following* P_1 *with high probability, specifically* $\lim_{\pi\to 0^+} P_e = 0$ *provided*

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

Proof: Employing sequential thresholding, the false positive event depends on the number of passes as $\alpha = 1/2^K$. With $K = \log_2\left(\frac{1-\pi}{\pi^{1+\delta}}\right)$ the probability the procedure returns a sequence corresponding to P_0 is then $P_e = \frac{\pi^\delta}{\pi^\delta + 1 - \beta}$, which, provided β is bounded away from one, goes to zero as π becomes small. As the maximum number of passes increases (as K increases), β also increases if m is fixed. To compensate, we must also increase m. The false negative probability is bounded away from one by a constant, and more strictly, tends to zero, provided $m > \frac{\log K}{D(P_0||P_1)}$, which follows by the Chernoff-Stein lemma for sufficiently large m [11]. We can bound the expected number of measurements as follows. First, consider the conditional expected number of measurements given P_0 is true, and correctly estimated:

$$\mathbb{E}[N_1 | x_1 = 0, \hat{x}_1 = 0] = \sum_{k=1}^{K} \frac{m}{2^{k-1}} \leq 2m$$

which follows the definition of the threshold. The remaining conditional expected number of samples can be bound by the

maximum number of passes, K, which gives

$$\begin{aligned}\mathbb{E}[N] &= \frac{\mathbb{E}[N_1]}{\alpha(1-\pi)+\pi(1-\beta)} \\ &\leq \frac{2m(1-\alpha)(1-\pi)+Km(\pi(1-\alpha)+\pi)}{\pi} \\ &\leq (1+\epsilon)\frac{2\log\log\left(\frac{1}{\pi}\right)}{\pi D(P_0||P_1)}\end{aligned}$$

which holds for all $\epsilon > 0$, and sufficiently small π. Thus, the procedure can drive the probability of error to zero provided $E[N] > 2\pi^{-1}D(P_0||P_1)^{-1}\log\log\left(\pi^{-1}\right)$. Lastly, the factor of two can be removed by redefining the initial threshold. We defer this detail, which is discussed in detail in [7]. ■

Remark 2. 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 $\log(1/\tilde{\pi}) \leq 1/\pi$.

VI. Fixed Sample Size Procedures

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

$$\mathbb{E}[N] = \frac{N_0}{\alpha(1-\pi)+\pi(1-\beta)}. \tag{21}$$

For comparison of the sampling requirements of non-sequential procedures with sequential procedures, we present a necessary condition for success. The theorem is again in terms of a lower bound on $\mathbb{E}[N]$ as a function of π.

Theorem 4. *Limitations of fixed sample size procedures. Any fixed sample size procedure with*

$$\mathbb{E}[N] < \frac{\log \pi^{-1}}{(1+\delta)\pi D(P_1||P_0)} \tag{22}$$

samples has $\lim_{\pi\to 0^+} P_e \geq \frac{\delta}{1+\delta}$.

Proof: We prove the theorem by contradiction. First, assume that $\lim_{\pi\to\infty} P_e < \frac{\delta}{1+\delta}$ and from (5) we have

$$\lim_{\pi\to 0^+} \frac{\alpha(1-\pi)}{\pi(1-\beta)} < \delta. \tag{23}$$

This implies $\alpha(1-\pi) < \delta\pi(1-\beta)$ for sufficiently small π. From (3),

$$\mathbb{E}[N] = \frac{N_0}{\alpha(1-\pi)+\pi(1-\beta)} > \frac{N_0}{(1+\delta)\pi(1-\beta)}. \tag{24}$$

From an application of the Chernoff-Stein Lemma [11], for any binary hypothesis test of sample size N_0, and fixed false negative probability, for sufficiently large N_0, any $\epsilon_1 > 0$,

$$\alpha > e^{-N_0 D(P_1||P_0)(1+\epsilon_1)} \tag{25}$$

and thus $N_0 > \frac{\log \alpha^{-1}}{(1+\epsilon_1)D(P_1||P_0)}$. This condition only holds for sufficiently large N_0, or since $N_0 \to \infty$ as $\alpha \to 0$, sufficiently small α. Combining this with (24) gives

$$\mathbb{E}[N] > \frac{\log \alpha^{-1}}{(1+\epsilon_1)D(P_1||P_0)(1+\delta)\pi(1-\beta)} \tag{26}$$

for all $\epsilon_1 > 0$. Lastly, from our original condition in (23), we have that $\alpha < \delta\pi(1-\beta)/(1-\pi)$ and thus

$$\begin{aligned}\mathbb{E}[N] &> \frac{\log\left(\frac{1-\pi}{\delta\pi(1-\beta)}\right)}{(1+\epsilon_1)D(P_1||P_0)(1+\delta)\pi(1-\beta)} \\ &> \frac{\log\left(\frac{1}{\pi}\right)+\log\left(\frac{1-\pi}{\delta(1-\beta)}\right)}{(1+\epsilon_1)D(P_1||P_0)(1+\delta)\pi}\end{aligned}$$

In the limit as $\pi \to 0$, we can overwhelm any lower order terms in π, and conclude (17) implies,

$$\mathbb{E}[N] > \frac{(1-\epsilon_2)\log\left(\pi^{-1}\right)}{(1+\epsilon_1)(1+\delta)\pi D(P_1||P_0)}$$

for any $\epsilon_1, \epsilon_2 > 0$, sufficiently small π. Taking the converse of this gives (22), completing the proof. ■

VII. Conclusion

This paper explored the problem of finding an atypical sequence amongst a large number of typical sequences. The results presented here quantify the number of samples required to recover an atypical sequence with high probability as the atypical sequences themselves become increasingly rare. We also proposed a robust procedure that requires less prior knowledge about the distributions. The procedures analyzed in this paper are robust to uncertainty in the prior probability of the rare distribution.

References

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] M. Malloy and R. D. Nowak, "On the limits of sequential testing in high dimensions," *CoRR*, vol. abs/1105.4540, 2011.

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

[9] 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

[10] 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

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