Sequential Testing for Sparse Recovery

Matthew L. Malloy, Robert D. Nowak

Abstract

This paper studies sequential methods for recovery of sparse signals in high dimensions. When compared to fixed sample size procedures, in the sparse setting, sequential methods can result in a particularly large reduction in the number of samples needed for reliable signal support recovery. Starting with a lower bound, we show *any* sequential sampling procedure fails in the high dimensional limit provided the average number of measurements per dimension is less then $D(P_0||P_1)^{-1}\log s$, where s is the level of sparsity and $D(P_0||P_1)$ the Kullback-Leibler divergence between the underlying distributions. An extension of the Sequential Probability Ratio Test (SPRT) which requires complete knowledge of the underlying distributions is shown to achieve this bound. We introduce a simple procedure termed *Sequential Thresholding* which can be implemented with limited knowledge of the underlying distributions, and guarantees exact support recovery provided the average number of measurements per dimension grows faster than $D(P_0||P_1)^{-1}\log s$, achieving the lower bound. For comparison, we show any *non-sequential* procedure fails provided the number of measurements grows at a rate less than $D(P_1||P_0)^{-1}\log n$, where n is the total dimension of the problem.

I. Introduction

Signal support recovery in high dimensions is a fundamental problem arising in many aspects of science and engineering. The goal of the basic problem is to determine, based on noisy observations, a sparse set of elements that somehow differ from the others.

In this paper we study the following problem. Consider a support set $\mathcal{S} \subset \{1,\dots,n\}$ and a random variable Y_i such that

$$Y_i \sim \begin{cases} P_0 & i \notin \mathcal{S} \\ P_1 & i \in \mathcal{S} \end{cases} \qquad i = 1,\dots,n$$

where P_0 and P_1 are probability measures on $\mathcal{Y}$. The dimension of the problem, n, is large – perhaps thousands or millions or more – but the support set $\mathcal{S}$ is sparse in the sense that the number of elements following distribution P_1 is much less than the dimension, i.e., $|\mathcal{S}| = s \ll n$. The goal of the sparse recovery problem is to identify the set $\mathcal{S}$ from multiple independent realizations of the random variables $Y_1, Y_2, \dots, Y_n$.

This work was supported in part by AFOSR grant number FA9550-09-1-0140 and FA9550-09-1-0643. Portions of this work were presented at the Asilomar Conference on Signals, Systems, and Computers, November 2011 [1], and the International Symposium on Information Theory in Saint Petersburg, August 2011 [2]. The authors would like to thank both Cun-Hui Zhang and Jarvis Haupt for helpful discussions.

The conventional theoretical treatment of this problem assumes that the samples are collected prior to data analysis in what is refereed to as a *non-sequential* (or *fixed sample size*) setting. In this case, m samples of each component are made (m samples of Y_i are gathered for each index i) and any test for inclusion in $\mathcal{S}$ is performed after the data is collected. The fundamental limits of reliable recovery are readily characterized in terms of Kullback-Leibler divergence and dimension (see Sec. III-B).

On the other hand, information gathering systems encountered in practice are often tasked with measuring some temporal signal or process, leaving the potential for the system to *adapt* the sampling approach based on prior observations. In this *sequential* setting, the decision to take an additional sample of any component i is based on prior realizations of that component. Herein lies the advantage of sequential methods: if prior samples indicate a particular component belongs (or doesn't belong) to $\mathcal{S}$ with sufficient certainty, measurement of that component can cease, and resources can be diverted to a more uncertain element. The focus of this paper is on the fundamental limits of recovery of such sequential systems.

A. Main Contributions

The results presented in this paper are in terms of asymptotic rate at which the average number of samples per dimension, denoted m, must increase with n to ensure exact recovery of $\mathcal{S}$ for any fixed distributions P_0 and P_1. For a given procedure, the probability of correctly recovering the set $\mathcal{S}$ depends on the triple (n, s, m). As the dimension of the problems grows (as $n \to \infty$), correctly recovering $\mathcal{S}$ becomes increasingly difficult, and m must also increase if we hope to recover $\mathcal{S}$. One manner in which we can quantify the performance of a procedure is the rate at which m must grow as a function of n and s to ensure recovery of $\mathcal{S}$.

As such, the main contributions are *1)* to derive a lower bound on the number of measurements required for success of any sequential procedure in the sparse setting, *2)* introduce a simple sequential procedure termed *Sequential Thresholding* which can be implemented with limited knowledge of the distributions and show this simple procedure is asymptotically optimal, *3)* compare this procedure to the known optimal SPRT which requires full knowledge of P_0 and P_1, and lastly *4)* compare these results to the performance of any non-sequential procedure. Table I summarizes these results.

TABLE I

AVERAGE NUMBER OF MEASUREMENTS PER DIMENSION FOR EXACT SUPPORT RECOVERY IN HIGH DIMENSIONAL LIMIT

any non-sequential	$m \geq \frac{\log n}{D(P_1 \| P_0)}$	necessary	
any sequential	$m \geq \frac{\log s}{D(P_0 \| P_1)}$	necessary	
SPRT based procedure	$m > \frac{\log s}{D(P_0 \| P_1)}$	sufficient	requires exact knowledge of P_0, P_1
Sequential Thresholding	$m > \frac{\log s}{D(P_0 \| P_1)}$	sufficient	can be implemented without exact knowledge of P_1

These developments are intriguing primarily for two reasons. First, *sequential* procedures can succeed when

the number of measurements per dimension increases at a rate logarithmic in the level of *sparsity*, i.e. $\log s$. In contrast, well known results from statistical testing show *non-sequential* procedures require the average number of measurements per dimension to increase at a rate logarithmic in the *dimension*, i.e. $\log n$. Secondly, and perhaps equally as surprising, *Sequential Thresholding*, a simple, practical procedure introduced here, achieves optimal performance as the dimension grows large.

B. Motivation

The problem of sparse signal recovery using sequential measurements arises in a number of commonly encountered problems in science and engineering. In communications, spectrum sensing for cognitive radio aims to identify unoccupied communication bands in the electromagnetic spectrum. Most bands will be occupied by primary users, but these users may come and go, leaving certain bands momentarily open and available for use by secondary transmitters. As the noisy samples of these occupied and un-occupied bands are collected in a temporal manner, sequential methods are a natural fit to map the occupation of the spectrum; in fact, recent work in spectrum sensing has given considerable attention to such approaches (see, for example [3], [4]).

Another captivating example a of sparse recovery problem where sequential methods are currently employed is that of the *Search for Extraterrestrial Intelligence* (SETI) project. Researchers at the SETI institute sense for narrowband electromagnetic energy from distant star systems using large antenna arrays, with the hopes of finding extraterrestrial transmission. The dimension of the problem consists of over 100 billion stars in the Milky Way alone, each with 9 million potential 'frequencies' in which to sense for narrow band energy. The subset of planetary systems with extraterrestrial transmission is sparse (as, of course, SETI is yet to make a contact). Moreover, while researchers may have a good idea of the distribution of the background noise, P_0, complete knowledge of P_1 is of course not available, making procedures based on sequential probability ratio testing impossible to implement. Roughly speaking, researchers at SETI use a sequential procedure that repeatedly tests energy levels against a threshold up to five times [5], [6]. If any of the up to five measurements are below the threshold, the procedure passes to the next frequency/star. Should the measurements exceed the threshold on all five occasions, measurements of that star and frequency are passed to an operator for further inspection. This procedure is closely related to Sequential Thresholding. Sequential Thresholding results in substantial gains over fixed sample size procedures and, unlike the SPRT, it can be computed without full knowledge of P_1.

Sparse recovery also underlies a number of recent micro-array studies in biology. Here, biologists attempt to estimate a sparse set of genes or proteins that are critically involved in a certain process or function (such as virus replication). The biologists may have good estimates of the null distribution, P_0, but not of the alternative distribution, P_1, again making procedures based on the SPRT impractical. A number of recent publications have implemented various multi-stage (thus sequential) procedures [7]–[10] that operate without full knowledge of P_1. The proposed procedures in general aim to reduce the total dimension of the problem and then employ traditional recovery techniques. While a number of authors suspect such sequential methods result in increased sensitivity, the gains are not theoretically quantified.

C. Related Work

Many of the fundamental results in sequential analysis were developed by Wald, and formalized in his book, Sequential Analysis [11]. Most relevant to the developments here, the sequential probability ratio test (SPRT) was shown to be optimal in terms of minimizing the error probabilities and expected number of measurements for a simple binary hypothesis test. A number of issues arise, including loss of optimality, when exact knowledge of the distributions is unavailable. In this scenario, a number of procedures have been proposed for composite alternatives, including those by Wald [12], which usually incorporate a weighting over parametric families, or generalized likelihood ratio testing [13].

Sequential testing for sparse signals was perhaps first studied by Posner in [14]. Motivated by the the problem of finding a lost satellite in the sky, Posner aimed to minimize the expected search time using a multistage procedure. Posner's procedure is closely related to the high dimensional extension of the sequential probability ratio test (see Sec. IV for details). Sequential approaches to the high dimensional sparse recovery problem have recently been given increased attention, perhaps motivated by the success of exploiting sparsity in other areas. Related work includes [15], [16], in which the authors extend the work of [14] to include multiple targets, encompassing a more general model.

In some of the first work to theoretically quantify the gains of sequential methods for high dimensional recovery [17], [18], the authors proposed a sequential procedure for recovery in additive Gaussian noise, termed *Distilled Sensing*. Our *Sequential Thresholding* approach is similar to the so-called Distilled Sensing method, however there are two main distinctions. First, the results in this paper are applicable to a larger class of problems characterized by finite Kullback-Leibler divergence; the Distilled Sensing approach is specific to the Gaussian setting. Second, here we are concerned with the probability of error in exact recover of $\mathcal{S}$; Distilled Sensing controls the false discovery and non-discovery rates which is less demanding than the error rate control. Also closely related to the lower bounds developed here are those of [19], which are in terms of the expected set difference, are restricted to the Gaussian setting, and published after the initial work in [1], [2].

Another related set of problems is that of finding the best arm in a muli-armed bandit game [20], [21]. Some approaches to these problems are similar in nature to Sequential Thresholding, namely the *median elimination* procedure of [22], but the problem of finding the best arm is fundamentally different than recovering $\mathcal{S}$. Another difference is these procedures in general assume no knowledge of the distributions P_0 and P_1, resulting in order optimal procedures at best.

D. Organization

The remainder of the paper is organized as follows. In Sec. II we formalize the problem. Sec. III-A derives the necessary condition on the number of samples required for exact recovery using any procedure. For comparison, Sec. III-B derives a neccessary condition on the average number of measurements for non-sequential procedures. Next, Sec. IV analyzes the SPRT in the sparse setting and discusses some of the shortcomings of the test when exact

knowledge of the distributions is not available. Lastly, Sec. V introduces *Sequential Thresholding* and analyzes its performance.

II. Problem Formulation

Let $\mathcal{S}$ be a subset of $\{1,\dots,n\}$ with cardinality $s=|\mathcal{S}|$. For any index $i\in\{1,\dots,n\}$, the random variable $Y_{i,j}$ is independent, identically distributed according to

$$Y_{i,j}\sim\begin{cases} P_0 & i\notin\mathcal{S} \\ P_1 & i\in\mathcal{S}\end{cases}\qquad i=1,\dots,n \tag{1}$$

where P_0 and P_1 are probability measures with joint support on $\mathcal{Y}$. In words, the random variable $Y_{i,j}$ follows distribution $P_1(\cdot)$ if i belongs to $\mathcal{S}$, and follows $P_0(\cdot)$ otherwise, and j indexes multiple i.i.d. samples of any component i. We refer to P_0 as the null distribution, and P_1 the alternative.

Our analysis is concerned with exact recovery of the set $\mathcal{S}$. Let $\hat{\mathcal{S}}$ be an estimate of $\mathcal{S}$. The family wise error rate is defined as:

$$\mathbb{P}_e=\mathbb{P}(\hat{\mathcal{S}}\neq\mathcal{S})=\mathbb{P}\left(\bigcup_{i\notin\mathcal{S}}\mathcal{E}_i\cup\bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right) \tag{2}$$

where $\mathcal{E}_i$ is the event that an error is made at index i.

In the lower bounds developed in this paper our consideration is limited to *component wise* procedures that test each index in an identical manner. This assumption implies that the individual error rates at each index are the same; specifically, $\mathbb{P}(\mathcal{E}_i)=\mathbb{P}(\mathcal{E}_{i'})$ for all $i,i'\in\mathcal{S}$, and, likewise $\mathbb{P}(\mathcal{E}_i)=\mathbb{P}(\mathcal{E}_{i'})$ for all $i,i'\notin\mathcal{S}$. Under this assumption we can simply notation and define the false positive and false negative rates:

$$\alpha=\mathbb{P}_0(\mathcal{E})\qquad\beta=\mathbb{P}_1(\mathcal{E})$$

where $\mathbb{P}_0(\mathcal{E})=\mathbb{P}(\mathcal{E}_i|i\notin\mathcal{S})$ and $\mathbb{P}_1(\mathcal{E})=\mathbb{P}(\mathcal{E}_i|i\in\mathcal{S})$. The component wise assumption implies the procedure only uses samples of component i to make inference about that particular component. More specifically, the decision to re-measure a particular component or include it in the estimate of $\mathcal{S}$ depends only on samples of that component. As the dimension of the problem grows large (which is our regime of interest), there is no loss of optimality associated with this restriction [19].

The log-likelihood ratio statistic comprised of multiple i.i.d. samples of a particular index is defined as:

$$L_i^{(\ell)}\left(Y_{i,1},\dots,Y_{i,\ell}\right):=\sum_{j=1}^{\ell}\log\frac{P_1(Y_{i,j})}{P_0(Y_{i,j})}. \tag{3}$$

Here, the superscript ℓ explicitly indicates the number of samples used to form the likelihood ratio and is suppressed when unambiguous. The Kullback-Leibler divergence from distribution P_1 to P_0 is defined as:

$$D(P_1||P_0)=\mathbb{E}_1\left[\log\frac{P_1(Y)}{P_0(Y)}\right]$$

where $\mathbb{E}_1[\cdot]$ is expectation with respect to distribution P_1, which gives the usual convergence of the normalized likelihood ratio as ℓ grows large:

$$\frac{1}{\ell} L_i^{(\ell)} \xrightarrow{a.s.} \begin{cases} -D(P_0||P_1) & i \notin \mathcal{S} \\ D(P_1||P_0) & i \in \mathcal{S}. \end{cases} \tag{4}$$

It is sometimes convenient to state results in terms of the maximum of $D(P_0||P_1)$ and $D(P_1||P_0)$. In this case, we define

$$D_{\mathrm{KL}} = \max\left\{D(P_0||P_1), D(P_1||P_0)\right\}.$$

In order to bound rates of convergence of particular testing procedures, we make use of the variance of the likelihood ratio, denoted

$$\sigma^2(P_1||P_0) = \mathrm{var}\left(L_i^{(1)} | i \in S\right) = \mathbb{E}_1\left[\left(\log\frac{P_1(Y)}{P_0(Y)} - D(P_1||P_0)\right)^2\right].$$

A sampling procedure Γ is a method used to determine the number of samples taken of each index. To be precise in characterizing a sampling procedure, we present three definitions.

Definition 1. Sampling procedure. *A binary valued function Γ on $(i,j) \in \{1,\ldots,n\} \times \mathbb{N}$ that defines the number of samples of Y_i that are observed. Specifically, if $\Gamma_{i,j} = 1$, then $Y_{i,j}$ is observed, and can be used in estimation of $\mathcal{S}$. Conversely, if $\Gamma_{i,j} = 0$, then $Y_{i,j}$ is not observed, and is not used in estimation of $\mathcal{S}$.*

Definition 2. Non-sequential (fixed sample size) sampling procedure. *Any sampling procedure Γ such that $\Gamma_{i,j}$ is not a function of $Y_{i',j'}$ for any i', j'.*

Definition 3. Sequential sampling procedure. *A sampling procedure Γ in which $\Gamma_{i,j}$ is allowed to depend on previous samples, specifically, $\Gamma_{i,j} : \{Y_{i,1},\ldots,Y_{i,j-1}\} \mapsto \{0,1\}$.*

Sequential procedures can make use of information as it becomes available to adjust the sample size, while non-sequential procedures, or *fixed sample size* procedures, fix the number of samples taken *a priori*. Note that under this definition, the set of non-sequential procedures are a subset of sequential procedures.

In order to make fair comparison between different procedures, we limit the total number of samples in expectation. For any procedure Γ we require

$$\mathbb{E}\left[\sum_{i,j} \Gamma_{i,j}\right] \leq nm \tag{5}$$

for some $m \geq 0$. This simply implies, on average, the procedure uses m or fewer samples per dimension.

The family wise error rate of any procedure used to estimate $\mathcal{S}$ depends on the underlying distributions P_0 and P_1, the dimension, n, the level of sparsity s, and the average number of samples per component, m. Throughout, s and m are non-decreasing functions in n (and thus, the set $\mathcal{S}$ is also a function of n). We suppress this dependence on n for ease of exposition. Our focus will be on finding the relationship between the triple (n, s, m) such that

for any fixed distributions P_0 and P_1, either $\lim_{n\to\infty} \mathbb{P}_e = 0$ (the procedure is reliable) or $\lim_{n\to\infty} \mathbb{P}_e > 0$ (the procedure is unreliable). Without loss of generality, we assume $s < n/2$ (since, of course, if $s \geq n/2$, one can re-label the problem, swapping P_0 and P_1). As we are interested in sparse problems, a few of the results require the assumption that $s \ll n$. We often make the assumption that $\lim_{n\to\infty} \frac{s}{n} = 0$, which is termed *sub-linear* sparsity, but this scaling is stated explicitly when needed.

III. Limits of Reliable Recovery

This section presents lower bounds on the number of measurements required for reliable recovery by any procedure in both the sequential and non-sequential setting. The bounds are in terms of the *expected* number of samples per dimension.

A. Limitation of Sequential Procedures

The following theorem quantifies the limitations of *any* procedure, which includes both sequential and non-sequential procedures, as non-sequential procedures are a subset of sequential procedures (from Def. 2 and Def. 3). The bound applies to finite problems, but also implies a necessary rate of growth as n becomes large, captured in the ensuing corollary.

Theorem 4. Finite sample limitations of sequential procedures. *Any (sequential) procedure with*

$$m \leq \frac{\log s + \log\left(\frac{1}{4\delta}\right)}{D_{\mathrm{KL}}}$$

also has

$$\mathbb{P}_e \quad \geq \quad 1 - e^{-\delta} \approx \delta$$

where the approximation holds for small δ.

Proof: See Appendix A. ∎

Thm. 4 establishes a lower bound on the expected number of samples needed to achieve a particular family wise error rate. As the dimension of the problem grows, it provides us with a necessary condition for reliable recovery.

Corollary 5. Limitations of sequential procedures. *Assume* $\lim_{n\to\infty} s/n = 0$. *Any (sequential) procedure with*

$$\lim_{n\to\infty} \frac{m}{\log s} \leq \frac{1}{D(P_0||P_1)}$$

also has $\lim_{n\to\infty} \mathbb{P}_e > 1/5$.

Proof: Thm. 4 implies that if $m \leq \frac{\log s}{D_{\mathrm{KL}}}$ then $\mathbb{P}_e \geq 1 - e^{-1/4} > 1/5$. Dividing by $\log s$ and taking the limit as $n \to \infty$ would give the lemma if $D_{\mathrm{KL}} = D(P_0||P_1)$. Instead, returning to the proof of Thm. 4, it is easily verified that if $\lim_{n\to\infty} s/n = 0$, the analysis follows with D_{KL} replaced by $D(P_0||P_1)$. ∎

In words, if the number of samples per dimension grows at a rate slower than logarithmically in the level of sparsity, no procedure can reliably recover $\mathcal{S}$. In shorthand notation, if $m \leq \frac{\log s}{D(P_0||P_1)}$ then $\mathbb{P}_e$ can not be driven to zero, and recovery of $\mathcal{S}$ is unreliable in the large n limit.

B. Limitation of Non-Sequential Procedures

Non-sequential methods, which sample each index a fixed number of times, can require significantly more measurements than sequential procedures. In the following theorem we state a necessary condition on m for any reliable non-sequential procedure. The proof is based on analysis of the *Chernoff Information* [23]. Our consideration is restricted to non-sequential procedures which sample each component $i = 1, \ldots, n$ exactly m times:

$$\Gamma_{i,j} = \begin{cases} 1 & j \leq m \\ 0 & j > m. \end{cases}$$

Theorem 6. Limitation of non-sequential procedures. *Assume* $\lim_{n\to\infty} s/n = 0$. *Any non-sequential procedure with*

$$\lim_{n\to\infty} \frac{m}{\log n} < \frac{1}{D(P_1||P_0)}. \tag{6}$$

also has

$$\lim_{n\to\infty} \mathbb{P}_e > 0.$$

Proof: See Appendix B. ■

IV. Sequential Probability Ratio Testing

A. The SPRT

Provided P_0 and P_1 are known, sequential probability ratio tests are optimal for binary hypothesis tests in terms of minimizing the expected number of measurements for any error probabilities α and β (shown originally in [24]); this optimality translates to the high dimensional case by simply considering n parallel SPRTs.

Each individual SPRT operates by continuing to measure a component if the corresponding likelihood ratio is within an upper and lower threshold, and terminating measurement otherwise. For scalar thresholds γ_{L} and γ_{U}, the procedure is defined as

$$\Gamma_{i,j'+1} \quad = \quad \begin{cases} 1 & \text{if} \quad \gamma_{\mathrm{L}} \leq \prod_{j=1}^{j'} \frac{P_1(Y_{i,j})}{P_0(Y_{i,j})} \leq \gamma_{\mathrm{U}} \\ 0 & \text{else} \end{cases} \tag{7}$$

where $\prod_{j=1}^{j'} \frac{P_1(Y_{i,j})}{P_0(Y_{i,j})}$ is the likelihood ratio comprised of all prior samples. If the likelihood ratio falls below γ_{L}, the SPRT labels index i as *not* belonging to $\hat{\mathcal{S}}$; if the likelihood ratio exceeds γ_{U}, index i is assigned to $\hat{\mathcal{S}}$. Equivalently, the test can be implement in the log-likelihood domain, and $L_i^{(j')}$ can be compared against $\log(\gamma_{\mathrm{L}})$ and $\log(\gamma_{\mathrm{U}})$. The procedure requires a random number of samples of each component, denoted J_i, and defined as

$$J_i := \min\{j : \Gamma_{i,j+1} = 0\}.$$

As we proceed we make a minor assumption on the distribution of the log-likelihood statistic. Specifically, the ensuing theorem and proof require existence of positive constants C_1 and C_2 such that

$$\mathbb{E}[L_i^{(J_i)} | L_i^{(J_i)} < \log\gamma_{\mathrm{L}}] \geq \log\gamma_{\mathrm{L}} - C_1 \qquad \mathbb{E}[L_i^{(J_i)} | L_i^{(J_i)} > \log\gamma_{\mathrm{U}}] \leq \log\gamma_{\mathrm{U}} + C_2 \tag{8}$$

for all thresholds γ_{U} and γ_{L}. In some cases, bounds for C_1 and C_2 are known (see [12], p.145, where explicit expressions for the Bernoulli and Gaussian case are given). In words, the requirement is the existence of a constant that bounds the expected value of the $\log$-likelihood ratio when the procedure terminates, regardless of the value of the threshold. This condition is satisfied when $L_i^{(1)}$ 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 [25].

Theorem 7. Ability of the SPRT. *The SPRT procedure with thresholds* $\gamma_{\mathrm{L}} = \frac{1}{s^{1+\epsilon}}$ *and* $\gamma_{\mathrm{U}} = (n-s)^{1+\epsilon}$*, any* $\epsilon > 0$*, has*

$$\lim_{n\to\infty} \mathbb{P}_e = 0$$

and

$$\lim_{n\to\infty} \frac{m}{\log s} \leq \frac{1+\epsilon}{D(P_0||P_1)}.$$

provided $s < n/\log n$*, and the condition in (8) is satisfied.*

Proof: See Appendix C. ∎

B. Implementation Issues

Implementing an SPRT on each component can be challenging for many problems encountered in practice. While the SPRT is optimal when both P_0 and P_1 are known and testing a single component 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. To see this more concretely, consider a problem where P_1 is a normal distribution with an unknown positive mean μ, and P_0 is a zero mean standard normal distribution. Here, the SPRT procedure continues to sample a particular index if

$$\frac{\log \gamma_{\mathrm{L}}}{j'\mu} + \frac{\mu}{2} \leq \frac{1}{j'} \sum_{j=1}^{j'} Y_{i,j} \leq \frac{\log \gamma_{\mathrm{U}}}{j'\mu} + \frac{\mu}{2}, \tag{9}$$

equivalent to (7). While the statistic $\sum_{j=1}^{j'} Y_{i,j}$ does not depend on the unknown parameter μ, the thresholds do. If the test is implemented with an incorrect value of μ, it may continue to sample an index *ad infinitum* (the so-called *open continuation region* [26]). This occurs when μ is overestimated; consider a scenario in which the threshold is set using $\tilde{\mu} = 10\mu$, where μ is the true mean of P_1. If P_1 is the true distribution, we have $\frac{1}{j'}\sum_{j=1}^{j'} Y_{i,j} \sim \mathcal{N}(\mu, 1/j')$, which, with high probability, will not exceed the upper threshold (which is greater than 5μ).

V. Sequential Thresholding

Sequential Thresholding is based on simple idea: repeatedly reduce the dimension of the problem by sequentially eliminating elements that exhibit strong evidence they don't belong to $\mathcal{S}$. Sequential Thresholding consists of a

series of K measurement steps, where each step eliminates from consideration a proportion of the components measured on the prior step. After the last step, the procedure terminates, and the remaining components are taken as the estimate of $\mathcal{S}$.

To illustrate the main idea behind the procedure, we first introduce a simplified version of Sequential Thresholding and analyze the simplified procedure for a specific problem. This *simple* Sequential Thresholding, while not achieving asymptotic optimality, does admit a simple error analysis. The more general version of Sequential Thresholding, which does achieve optimality and the lower bound of Cor. 5, is presented in the second half of this section.

A. Example of Simple Sequential Thresholding

To highlight the main idea behind Sequential Thresholding, and the potential performance gains, consider a problem where $P_0 \sim \mathcal{N}(0,1)$ and $P_1 \sim \mathcal{N}(\theta,1)$ for some $\theta > 0$. The *simple* Sequential Thresholding procedure requires two inputs - *1)* the number of steps, $K \approx \log n$, and *2)* an even integer $m \geq 2$ that defines the average number of samples per index, and hence the total budget. On the first step the procedure samples all indices $m/2$ times each, for all i, requiring $mn/2$ samples. These $m/2$ samples are summed for each index. If this sum, $\sum_{j=1}^{m/2} Y_i$, is less than zero, that particular index is not sampled on subsequent passes. This eliminates a proportion (approximately half) of components following the null distribution (since the median of $\sum_{j=1}^{m/2} Y_i$ for $i \notin S$ is zero). Indices that exceed the threshold, i.e. $\{i : \sum_{j=1}^{m/2} Y_i > 0\}$, are sampled on the subsequent step. This process continues for $K \approx \log_2 n$ steps. After the Kth step, the procedure terminates, and estimates $\mathcal{S}$ as the set of indices that have not been eliminated from consideration. Roughly speaking, provided $s \ll n$, the procedure reduces the number of samples taken on each step by half as most components follow the null, which is zero mean. The total number of samples required by the procedure on all passes is $mn/2 + mn/4 + mn/8 + \cdots \approx mn$. This implies, on average, m or fewer samples per dimension.

Algorithm 1 Simple Implementation of Sequential Thresholding

input: $K \approx \log n$ steps, budget $m \geq 2$

initialize: $\mathcal{S}_1 = \{1, \ldots, n\}$

for $k = 1, \ldots, K$ **do**

  for $i \in \mathcal{S}_k$ **do**

    measure: $\{Y_{i,j}\}_{j=1}^{m/2} \sim \begin{cases} \prod_{j=1}^{m/2} P_0(Y_{i,j}) & i \notin \mathcal{S} \\ \prod_{j=1}^{m/2} P_1(Y_{i,j}) & i \in \mathcal{S} \end{cases}$

    threshold: $\mathcal{S}_{k+1} := \{i \in \mathcal{S}_k : \sum_{j=1}^{m/2} Y_{i,j} > 0\}$

  end for

end for

output: $\mathcal{S}_{K+1}$

Corollary 8. Ability of simple Sequential Thresholding. *Let* $K = \lceil (1+\epsilon)\log_2 n \rceil$ *for any* $\epsilon > 0$ *and consider*

the setting above where $P_0 \sim \mathcal{N}(0,1)$ *and* $P_1 \sim \mathcal{N}(\theta,1)$. *The simple Sequential Thresholding algorithm satisfies* $\lim_{n\to 0} \mathbb{P}_e = 0$ *provided*

$$m > \frac{\log s + \log\log_2 n}{\theta^2/8}.$$

Proof: From a union bound,

$$\mathbb{P}_e \leq (n-s)\alpha + s\beta. \tag{10}$$

The false positive event occurs when, for $i \notin \mathcal{S}$, the index survives all K thresholding steps. By the independence across steps, and since the median of $\sum_{j=1}^{m/2} Y_i$ for $i \in \mathcal{S}$ is zero,

$$\alpha = \left(\frac{1}{2}\right)^K \leq \left(\frac{1}{n}\right)^{(1+\epsilon)}. \tag{11}$$

The false negative event occurs when for some $i \in \mathcal{S}$, the sum $\sum_{j=1}^{m/2} Y_i$ falls below zero on any step. Applying a union bound and Gaussian tail bound, since $\sum_{j=1}^{m/2} Y_i \sim \mathcal{N}\left(m\theta/2, m/2\right)$, we have

$$\begin{aligned}
\beta &\leq \frac{K}{2}\exp\left(-m\theta^2/4\right) \\
&\leq (1+\epsilon)\log_2(n)\exp\left(-m\theta^2/4\right) \\
&\leq \exp\left(-m\theta^2/4 + \log\left((1+\epsilon)\log_2 n\right)\right).
\end{aligned} \tag{12}$$

Combining (10), (11) and (12) gives

$$\mathbb{P}_e \leq (n-s)\left(\frac{1}{n}\right)^{(1+\epsilon)} + \exp\left(-m\theta^2/4 + \log s + \log\left((1+\epsilon)\log_2 n\right)\right).$$

Imposing the condition in the theorem, and taking the limit as $n \to \infty$ gives the desired result. ■

While the sub-optimal simple version of Sequential Thresholding does not achieve the lower bound, it does outperform non-sequential procedures. The procedure requires the average number of samples per dimension, m, to be order $\log s + \log\log n$ for successful recovery. On the other hand, Sec. III-B shows non-sequential methods require m on the order of $\log n$ samples. For large n and small s, $\log n$ can be significantly larger than $\log s + \log\log n$, implying that Sequential Thresholding, for sufficiently sparse problems, will succeed with fewer samples than any non-sequential procedure.

B. Details of Sequential Thresholding

While the previous discussion highlighted the main principle behind Sequential Thresholding, the procedure becomes slightly more complicated in its full generality. To the show procedure achieves the lower bound of Cor. 5 as n grows large, both the allocation of measurements across steps and the proportion of null components discarded on each step must be adjusted.

In general, Sequential Thresholding requires three inputs: *1)* K, the number of steps, *2)* a constant $\rho \in [1/2, 1)$ representing the proportion of null components discarded on each step, and *3)* a total measurement budget mn

(implying m samples per dimension). The expected proportion of null components discarded on each step, ρ, is fixed throughout the procedure and defined as

$$\mathbb{P}_0(L_i^{(m_k)} \leq \gamma_k) = \rho. \tag{13}$$

Here, m_k is the number of samples used on any particular index measured on step k. As m_k is a function of the step index, so is the threshold γ_k.

With ρ and K as inputs, and a total expected measurement budget mn, Sequential Thresholding operates as follows. Let $\mathcal{S}_k$ denote the subset of $\{1, \ldots, n\}$ comprised of components still under consideration at step k. The procedure first initializes by setting $\mathcal{S}_1 = \{1, \ldots, n\}$. For steps $k = 1, \ldots, K$, the procedure proceeds as follows. On step k, each component in $\mathcal{S}_k$ is sampled m_k times. The number of samples taken on step k is defined as

$$m_k = \left\lfloor m\, k\, \rho^2 \left(\frac{n}{n + sK^2} \right) \right\rfloor. \tag{14}$$

The procedure then compares the likelihood ratio comprised of the m_k samples to the threshold defined in (13) and includes only the indices that exceed the threshold in the set of components to be sampled on the following step:

$$\mathcal{S}_{k+1} = \left\{ i : L_i^{(m_k)} > \gamma_k, i \in \mathcal{S}_k \right\}$$

where γ_k is defined in (13). In words, if $L_i^{(m_k)}$ is below γ_k, no further measurements of component i are taken for the remainder of the procedure. Otherwise, component i is measured on the subsequent step. By definition of γ_k, approximately ρ times the number of remaining components following P_0 will be eliminated on each step; if $s \ll n$, each thresholding step eliminates approximately ρ times the total number of components remaining. After step K, the procedure terminates and estimates $\mathcal{S}$ as the indices still under consideration: $\hat{\mathcal{S}} = \mathcal{S}_{K+1}$. The procedure is detailed in Alg. 2.

Algorithm 2 Sequential Thresholding

input: $K = \left\lceil \log_{\frac{1}{1-\rho}} \left(\frac{2(n-s)}{\delta} \right) \right\rceil$ steps, $\rho \in [1/2, 1)$, maximum budget m

initialize: $\mathcal{S}_1 = \{1, \ldots, n\}$

for $k = 1, \ldots, K$ **do**

  for $i \in \mathcal{S}_k$ **do**

    measure: $\{Y_{i,j}^{(k)}\}_{j=1}^{m_k} \sim \begin{cases} \prod_{j=1}^{m_k} P_0(Y_{i,j}^{(k)}) & i \notin \mathcal{S} \\ \prod_{j=1}^{m_k} P_1(Y_{i,j}^{(k)}) & i \in \mathcal{S} \end{cases}$ where m_k is given in (14)

    threshold: $\mathcal{S}_{k+1} := \{i \in \mathcal{S}_k : L_{i,k}^{(m_k)} > \gamma_k\}$

  end for

end for

output: $\mathcal{S}_{K+1}$

C. Measurement Budget

By design, Sequential Thresholding satisfies the measurement budget in (5). Consider the expected number of samples required by Sequential Thresholding:

$$\begin{aligned}\mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\right] &= \mathbb{P}\left(\bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right)\mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\middle|\bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right] \\ &\quad + \mathbb{P}\left(\bigcap_{i\in\mathcal{S}}\mathcal{E}_i^c\right)\mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\middle|\bigcap_{i\in\mathcal{S}}\mathcal{E}_i^c\right] \end{aligned} \tag{15}$$

where the equality follows from the law of total probability and conditioning on one or more false negative events. From the description of the procedure, one or more false negatives can only reduce the total number of samples, and we have

$$\mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\middle|\bigcap_{i\in\mathcal{S}}\mathcal{E}_i^c\right] \geq \mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\middle|\bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right].$$

Combining this with (15) gives

$$\begin{aligned}\mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\right] &\leq \mathbb{E}\left[\sum_{i,j}\Gamma_{i,j}\middle|\bigcap_{i\in\mathcal{S}}\mathcal{E}_i^c\right] \\ &= \sum_{k=1}^{K} m_k\left((1-\rho)^{k-1}(n-s)+s\right) \\ &\leq \sum_{k=1}^{K} m\left(\frac{n}{n+sK^2}\right)k\rho^2\left((1-\rho)^{k-1}(n-s)+s\right) \\ &\leq mn\left(\frac{n-s+sK^2}{n+sK^2}\right) \\ &\leq mn.\end{aligned}$$

Here, the equality follows from independence of the samples across the K steps. The third inequality follows as the sum is a geometric series for $\rho \in [1/2, 1)$ and as $\sum_{k=1}^{K} k \leq K^2$.

D. Ability of Sequential Thresholding

For fixed P_0 and P_1, the following theorem and corollary relate (n, s, m) to the family wise error rate of the procedure. The corollary follows from the more general Thm. 10.

Corollary 9. Ability of Sequential Thresholding. *If*

$$\lim_{n\to\infty}\frac{m}{\log s} > \frac{1}{D(P_0||P_1)}$$

then sequential threshold satisfies

$$\lim_{n\to\infty}\mathbb{P}_e = 0$$

with input parameters $K = \lceil \log(2(n-s))/(2\log\log s) - 1/2 \rceil$ *and* $\rho = 1 - \frac{1}{\sqrt{\log s}}$ *provided* $s < n/(\log n)^2$ *and* $\lim_{n\to\infty} s = \infty$.

Proof: The proof follows from Thm. 10 by setting $\delta = \frac{1}{\log s}$, which implies $\lim_{n\to\infty} \mathbb{P}_e = 0$. By setting K and ρ as defined in the statement of the theorem, $\lim_{n\to\infty} c_n = D(P_0||P_1)$, where c_n is defined in (16). Together with the forward part of the theorem, this implies the corollary. ■

Comparison of Cor. 9 to Cor. 5 shows that Sequential Thresholding is asymptotically optimal in terms of the required number of samples needed for reliable recovery. We continue with the main theorem of Sequential Thresholding, which quantifies the expected number of samples per dimension in the finite setting. The theorem is in terms of a sequence, c_n, which, under certain conditions, approaches the Kullback-Leibler divergence between P_0 and P_1. Proof of the theorem relies on techniques closely related to the Chernoff-Stein Lemma, and is found in the Appendix.

Theorem 10. Finite sample performance of Sequential Thresholding. *Consider Sequential Thresholding with* $K = \left\lceil \log_{\frac{1}{1-\rho}} \left(\frac{2(n-s)}{\delta} \right) \right\rceil$ *steps and measurement allocation in (14). Provided*

$$m \geq \frac{\log s + \log \delta^{-1} + \log 4}{c_n}$$

then

$$\mathbb{P}_e \leq \delta$$

where

$$c_n = \rho^2 \left(\frac{n}{n + sK^2} \right) \left(D(P_0||P_1) - \sqrt{\frac{\sigma^2(P_0||P_1)}{\left(\frac{\rho^2 n \log s}{D(P_0||P_1)(n+sK^2)} - 1 \right)(1-\rho)}} \right) \tag{16}$$

and is assumed to be positive.

Proof: See Appendix D. ■

Thm. 10 and Cor. 9 imply that as the size of the problem increases (as n goes to infinity), if m is greater than $D(P_0||P_1)^{-1} \log s$, the procedure will succeed in exact recovery of the sparse support set. This achieves the lower bound in Cor. 5, which states that any reliable procedure requires at least $D(P_0||P_1)^{-1} \log s$ samples per dimension. This implies that Sequential Thresholding is in a sense *first order* optimal. While not discussed here, one could also analyze the rate at which the procedure approaches the lower bound in Cor. 5, although the authors suspect the procedure would not achieve *second* order optimality.

E. Implementation and Practical Concerns

One of the main attributes of Sequential Thresholding is that implementation does not require exact knowledge of distribution P_1. While apparent in the example of simple Sequential Thresholding at the beginning of the section, in which two normal distributions are compared, this appealing property also extends to other settings. To be more

specific, consider a sparse recovery problem in which P_0 is known (or well approximated), but P_1 is defined by a parametric family of distributions with an unknown parameter θ. Clearly the likelihood ratio,

$$L(Y) = \log\left(\frac{P_1(Y,\theta)}{P_0(Y)}\right)$$

cannot be formed as θ is unknown. None-the-less, if the likelihood ratio is a monotonic function of a test statistic T which does not depend on the unknown parameter θ, then an equivalent test based on T can be written; (13) can be written in terms of T. As the test statistic and the thresholds, γ_k, depend only on P_0, the procedure can be implemented without knowledge of θ.

In practice this occurs for many distributions in the exponential family as the log-likelihood ratio, L_i, defined in (3), is a monotonic function of a test statistic T that does not depend on parameters of P_1. This property is well illustrated by the example of testing two Gaussian distributions discussed in the beginning of the section. If we assume the null distribution is known, but the offset mean of P_1 is unknown, the procedure can still be implemented. The sum of the measurements, $\sum_j Y_{i,j}$, is a sufficient statistic whose distribution under P_0 does not depend on P_1. This implies that γ_k does not depend on P_1, and thus, the procedure can be implemented without this knowledge.

There are two possible implementations of Sequential Thresholding which we refer to as *parallel* and *scanning*. The parallel implementation samples and tests all n components in parallel according to the procedure. The scanning implementation measures and tests the n components in a sequence (which can be arbitrary). For example, the scanning implementation can begin with component $i = 1$ and repeatedly measure and threshold the observations up to K times. If an observation falls below the threshold at any point, then the scanning procedure immediately moves on to the next component. If K observations are made without an observation falling below the threshold, then the component is added to the set S_{K+1}.

The two implementations are equivalent from a theoretical perspective. The parallel implementation may be more natural for large-scale experimental designs (e.g., in the biological sciences), whereas the scanning implementation is more appropriate in communications applications such as spectrum sensing. The latter also reveals natural connections between Sequential Thresholding and the SPRT.

VI. Conclusion

This paper showed that sequential methods for support recovery of high dimensional sparse signals in noise can succeed using far fewer measurements than non-sequential methods. More specifically, non-sequential methods require the number of measurements to grow logarithmically with the dimension, while sequential methods succeed if the number of measurements grows logarithmically with the level of sparsity. A simple procedure termed Sequential Thresholding was shown to achieve the lower bound asymptotically. Sequential Thresholding can be implemented without full knowledge of the underlying distributions, making it more practical for sparse recovery encountered in real world problems.

REFERENCES

[1] 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*. IEEE, 2011, pp. 1245–1249.

[2] ——, "Sequential analysis in high-dimensional multiple testing and sparse recovery," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 2661–2665.

[3] A. Tajer, R. Castro, and X. Wang, "Adaptive spectrum sensing for agile cognitive radios," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2966 –2969.

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

[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*, January 19, 2012.

[7] H. Müller, R. Pahl, and H. Schäfer, "Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies," *Genetic epidemiology*, vol. 31, no. 8, pp. 844–852, 2007.

[8] J. M. Satagopan and R. C. Elston, "Optimal two-stage genotyping in population-based association studies," *Genetic Epidemiology*, vol. 25, no. 2, pp. 149–157, 2003.

[9] S. Zehetmayer, P. Bauer, and M. Posch, "Two-stage designs for experiments with a large number of hypotheses," *Bioinformatics*, vol. 21, pp. 3771–3777, October 2005.

[10] L. Hao, A. Sakurai, T. Watanabe, E. Sorensen, C. A. Nidom, M. A. Newton, P. Ahlquist, and Y. Kawaoka, "Drosophila RNAi screen identifies host genes important for influenza virus replication," *Nature*, Jul. 2008.

[11] A. Wald, *Sequential Analysis*. Hoboken, NJ, USA: John Wiley and Sons, Inc., 1947.

[12] ——, "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

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

[14] E. Posner, "Optimal search procedures," *Information Theory, IEEE Transactions on*, vol. 9, no. 3, pp. 157 – 160, jul 1963.

[15] E. Bashan, R. Raich, and A. Hero, "Optimal two-stage search for sparse targets using convex criteria," *Signal Processing, IEEE Transactions on*, vol. 56, no. 11, pp. 5389 –5402, nov. 2008.

[16] E. Bashan, G. Newstadt, and A. Hero, "Two-stage multiscale search for sparse targets," *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2331 –2341, may 2011.

[17] J. Haupt, R. Castro, and R. Nowak, "Adaptive discovery of sparse signals in noise," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, oct. 2008, pp. 1727 –1731.

[18] ——, "Distilled Sensing: Adaptive sampling for sparse detection and estimation," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6222–6235, 2011.

[19] R. Castro, "Adaptive sensing performance lower bounds for sparse signal estimation and testing," *arXiv preprint arXiv:1206.0648*, 2012.

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

[21] S. Mannor and J. N. Tsitsiklis, "The Sample Complexity of Exploration in the Multi-Armed Bandit Problem," *Journal of Machine Learning Research*, vol. 5, pp. 623–648, 2004.

[22] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *Computational Learning Theory*. Springer, 2002, pp. 193–209.

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

[24] 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.

[25] 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

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

APPENDIX A
PROOF OF THEOREM 4

Proof: We first bound the family wise error rate in terms of the false positive and false negative probabilities associated with incorrectly assigning or excluding any element from $\hat{\mathcal{S}}$. From (2),

$$\begin{aligned}\mathbb{P}_e &= \mathbb{P}\left(\bigcup_{i\notin\mathcal{S}}\mathcal{E}_i \cup \bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right) \\ &= 1-\mathbb{P}\left(\bigcap_{i\notin\mathcal{S}}\mathcal{E}_i^c \cap \bigcap_{i\in\mathcal{S}}\mathcal{E}_i^c\right).\end{aligned}$$

By the assumption of a coordinate wise test,

$$\begin{aligned}\mathbb{P}_e &= 1-(1-\beta)^s(1-\alpha)^{n-s} \\ &\geq 1-e^{-\beta s}e^{-\alpha(n-s)} \qquad (17)\end{aligned}$$

where the last inequality follows as $(1-\beta)^n \leq e^{-\beta n}$ for $\beta \in [0,1]$ and $n \in \{1,\dots\}$. To continue, we can bound the expected number of samples associated with any particular index. From [26], Thm. 2.39, the following holds for any binary hypothesis test

$$m_0 \geq \frac{\alpha\log\left(\frac{\alpha}{1-\beta}\right)+(1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right)}{D(P_0||P_1)}$$

where m_0 is the expected number of samples of any component $i \notin \mathcal{S}$. We can further bound the expected number of samples as

$$\begin{aligned}m_0 &\geq \frac{\alpha\log\alpha+(1-\alpha)\log(1-\alpha)+(1-\alpha)\log\beta^{-1}}{D(P_0||P_1)} \\ &\geq \frac{(1-\alpha)\log\beta^{-1}-\log 2}{D(P_0||P_1)}\end{aligned}$$

where the first inequality follows as $\alpha\log(1/(1-\beta)) \geq 0$, and the last inequality follows as $\alpha\log\alpha+(1-\alpha)\log(1-\alpha) \geq \log(1/2)$, all $\alpha \in [0,1]$. Likewise

$$\begin{aligned}m_1 &\geq \frac{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right)+\beta\log\left(\frac{\beta}{1-\alpha}\right)}{D(P_1||P_0)} \\ &\geq \frac{(1-\beta)\log\alpha^{-1}-\log 2}{D(P_1||P_0)}\end{aligned}$$

where m_1 is the average number of samples given $i \in \mathcal{S}$, and the first inequality is again from Thm. 2.39 of [26]. Let $D_{\mathrm{KL}} = \max\{D(P_0||P_1), D(P_1||P_0)\}$. We have

$$\begin{aligned}m &= \frac{(n-s)m_0+s\,m_1}{n} \\ &\geq \frac{(n-s)(1-\alpha)\log\beta^{-1}+s(1-\beta)\log\alpha^{-1}-n\log 2}{nD_{\mathrm{KL}}}.\end{aligned}$$

If $\alpha \leq \beta$ we have

$$\begin{aligned} m &\geq \frac{(n-s)(1-\beta)\log\beta^{-1} + s(1-\beta)\log\beta^{-1} - n\log 2}{nD_{\mathrm{KL}}} \\ &= \frac{(1-\beta)\log\beta^{-1} - \log 2}{D_{\mathrm{KL}}} \\ &\geq \frac{\log\left(\frac{1}{2\beta}\right) - \log 2}{D_{\mathrm{KL}}} \end{aligned}$$

where the last inequality is easily verified for $\beta \in [0,1]$.

Imposing the condition in the forward part of the theorem, $m \leq (\log s + \log(4\delta)^{-1})/D_{\mathrm{KL}}$ gives

$$\frac{\log s + \log\left(\frac{1}{4\delta}\right)}{D_{\mathrm{KL}}} \geq m \geq \frac{\log\left(\frac{1}{2\beta}\right) - \log 2}{D_{\mathrm{KL}}}$$

which implies

$$\log s + \log\left(\frac{1}{4\delta}\right) \geq \log\left(\frac{1}{2\beta}\right) - \log 2$$

and thus

$$\beta \geq \frac{\delta}{s}.$$

Hence,

$$\begin{aligned} \mathbb{P}_e &\geq 1 - e^{-\delta}e^{-(n-s)\alpha} \\ &\geq 1 - e^{-\delta}. \end{aligned}$$

Conversely if $\beta > \alpha$

$$m \geq \frac{\log\left(\frac{1}{2\alpha}\right) - \log 2}{D_{\mathrm{KL}}}$$

and

$$\mathbb{P}_e \geq 1 - e^{-s\beta}e^{-\frac{(n-s)\delta}{s}\alpha}$$

which, provided $s < n/2$, gives

$$\mathbb{P}_e \geq 1 - e^{-\delta}.$$

completing the proof. ■

APPENDIX B
PROOF OF THEOREM 6

Proof: We write the family wise error rate as:

$$\begin{aligned} \mathbb{P}_e &= \mathbb{P}\left(\bigcup_{i\notin\mathcal{S}}\mathcal{E}_i \cup \bigcup_{i\in\mathcal{S}}\mathcal{E}_i\right) \\ &= 1 - (1-\beta)^s(1-\alpha)^{n-s} \\ &\geq 1 - e^{-\alpha(n-s)}e^{-\beta s} \end{aligned} \tag{18}$$

which cannot be driven to zero if

$$\alpha > \frac{1}{n-s}. \tag{19}$$

Equivalently, if

$$\lim_{m\to\infty} \frac{1}{m} \log \alpha^{-1} < \lim_{m\to\infty} \frac{1}{m} \log(n-s)$$

then $\lim_{n\to\infty} \mathbb{P}_e > 0$.

From [23], p. 386, (Chernoff Information),

$$\lim_{m\to\infty} \frac{1}{m} \log \alpha^{-1} = D(P_\lambda || P_0) \leq D(P_1 || P_0) \tag{20}$$

where

$$P_\lambda = \frac{P_0^\lambda P_1^{1-\lambda}}{\int_\Omega P_0^\lambda P_1^{1-\lambda} dy}$$

for $\lambda \in [0,1]$. Thus, if

$$\lim_{m\to\infty} \frac{1}{m} \log(n-s) \geq D(P_1 || P_0)$$

$\lim_{n\to\infty} \mathbb{P}_e > 0$. Since $\lim_{n\to\infty} m = \infty$ and $\lim_{n\to\infty} n - s = n$, this implies the result. If

$$\lim_{n\to\infty} \frac{m}{\log n} < \frac{1}{D(P_1 || P_0)} \tag{21}$$

then $\lim_{n\to\infty} \mathbb{P}_e > 0$. ■

APPENDIX C
PROOF OF THEOREM 7

Proof: For an SPRT with thresholds γ_{L} and γ_{U}, from [26], the following well known inequalities hold:

$$\alpha \leq \gamma_{\mathrm{U}}^{-1} = \frac{1}{(n-s)^{1+\epsilon}} \qquad \beta \leq \gamma_{\mathrm{L}} = \frac{1}{s^{1+\epsilon}}. \tag{22}$$

From a union bound on the family-wise error rate

$$\lim_{n\to\infty} \mathbb{P}_e \leq \lim_{n\to\infty} (n-s)\alpha + s\beta = 0 \tag{23}$$

implying the forward portion of the lemma.

We can write the expected number of measurements per dimension as

$$m = \frac{(n-s)\mathbb{E}_0[J_i] + s\,\mathbb{E}_1[J_i]}{n}$$

By Wald's identity [26]

$$\mathbb{E}_1[J_i] = \frac{\mathbb{E}_1\left[L_i^{(J_i)}\right]}{\mathbb{E}_1\left[L_i^{(1)}\right]} = \frac{\mathbb{E}_1\left[L_i^{(J_i)}\right]}{D(P_1||P_0)}$$

and similarly, $\mathbb{E}_0\left[J_i\right] = \frac{-\mathbb{E}_0\left[L_i^{(J_i)}\right]}{D(P_0||P_1)}$. Dividing by $\log s$ and taking the limit, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{m}{\log s} &= \lim_{n\to\infty} \frac{(n-s)\mathbb{E}_0\left[J_i\right] + s\,\mathbb{E}_1\left[J_i\right]}{n\log s} \\
&= \lim_{n\to\infty} \frac{-(n-s)\mathbb{E}_0\left[L_i^{(J_i)}\right]}{n\log(s)D(P_0||P_1)} + \lim_{n\to\infty} \frac{s\,\mathbb{E}_1\left[L_i^{(J_i)}\right]}{n\log(s)D(P_1||P_0)} \\
&\leq \lim_{n\to\infty} \frac{(n-s)(\log\gamma_{\mathrm{L}}^{-1} + C_1)}{n\log(s)D(P_0||P_1)} + \lim_{n\to\infty} \frac{s(\log\gamma_{\mathrm{U}} + C_2)}{n\log(s)D(P_0||P_1)}
\end{aligned}$$

where the inequality follows by the assumptions in (8) and as

$$\begin{aligned}
-\mathbb{E}_0\left[L_i^{(J_i)}\right] &= -\left((1-\alpha)\mathbb{E}_0\left[L_i^{(J_i)}\middle|\, L_i^{(J_i)} < \log\gamma_{\mathrm{L}}\right] + \alpha\mathbb{E}_0\left[L_i^{(J_i)}\middle|\, L_i^{(J_i)} > \gamma_{\mathrm{U}}\right]\right) \\
&\leq (1-\alpha)(\log\gamma_{\mathrm{L}}^{-1} + C_1) + \alpha(\log\gamma_{\mathrm{U}}^{-1} - C_2) \\
&\leq (1-\alpha)(\log\gamma_{\mathrm{L}}^{-1} + C_1) \\
&\leq \log\gamma_{\mathrm{L}}^{-1} + C_1
\end{aligned}$$

and likewise,

$$\mathbb{E}_1\left[L_i^{(J_i)}\right] \leq \log\gamma_{\mathrm{U}} + C_2.$$

Using the prescribed values of γ_{U} and γ_{L} gives

$$\lim_{n\to\infty} \frac{m}{\log s} = \frac{1+\epsilon}{D(P_0||P_1)} \tag{24}$$

provided $s < n/\log n$, completing the proof. ■

APPENDIX D
PROOF OF THEOREM 10

Proof: From the union bound on the family wise error rate, we have

$$\mathbb{P}_e \leq (n-s)\alpha + s\beta. \tag{25}$$

The false negative event is given as

$$\begin{aligned}
\beta &= \mathbb{P}_1\left(\bigcup_{k=1}^{K} L_{i,k}^{(m_k)} < \gamma_k\right) \\
&\leq \sum_{k=1}^{K} \mathbb{P}_1\left(L_{i,k}^{(m_k)} < \gamma_k\right)
\end{aligned}$$

We continue by bounding the above probability. The following analysis is closely related to the Chernoff-Stien lemma [23], but modified for one sided tests. Let $\boldsymbol{y}_k = (y_1, \ldots, y_{m_k})$ and define the region $\mathcal{A}_k \subset \mathbb{R}^{m_k}$ as

$$\mathcal{A}_k := \{\boldsymbol{y}_k : L^{(m_k)}(\boldsymbol{y}_k) < \gamma_k\}.$$

For all $\boldsymbol{y}_k$ in $\mathcal{A}_k$, by definition,

$$L^{(m_k)}(\boldsymbol{y}_k) = \sum_{j=1}^{m_k} \log\frac{P_1(y_j)}{P_0(y_j)} < \gamma_k$$

which implies

$$P_1^{m_k}(\boldsymbol{y}_k) < e^{\gamma_k} P_0^{m_k}(\boldsymbol{y}_k)$$

where $P_1^{m_k}(\boldsymbol{y}_k) = \prod_{j=1}^{m_k} P_1(y_j)$. Again by definition (suppressing subscript k for ease of notation)

$$\begin{aligned}
\mathbb{P}_1\left(L^{(m_k)} < \gamma_k\right) &= \int_{\boldsymbol{y}\in\mathcal{A}} P_1^{m_k}(\boldsymbol{y})d\boldsymbol{y} \\
&\leq \int_{\boldsymbol{y}\in\mathcal{A}} e^{\gamma_k} P_0^{\rho m}(\boldsymbol{y})d\boldsymbol{y} \\
&= e^{\gamma_k}\int_{\boldsymbol{y}\in\mathcal{A}} P_0^{\rho m}(\boldsymbol{y})d\boldsymbol{y} \\
&\leq e^{\gamma_k}. \qquad (26)
\end{aligned}$$

The above relationship holds for any γ_k, though only proves meaningful if set correctly. To this end, for some $\epsilon_k > 0$, let

$$\gamma_k = -m_k(D(P_0||P_1) - \epsilon_k) \qquad (27)$$

which from (26) gives

$$\mathbb{P}_1\left(L^{(m_k)} < \gamma\right) \leq e^{-m_k(D(P_0||P_1)-\epsilon_k)}.$$

It remains to show what values of ϵ_k simultaneously satisfy (13) for any $\rho \in [1/2, 1)$. Specifically, we need to find the range values of ϵ_k that satisfy

$$\mathbb{P}_0\left(L^{(m_k)} \leq -m_k(D(P_0||P_1) - \epsilon_k)\right) \geq \rho.$$

Proceeding,

$$\begin{aligned}
\mathbb{P}_0\left(L^{(m_k)} \leq -m_k(D(P_0||P_1) - \epsilon_k)\right) &= \mathbb{P}_0\left(\frac{1}{m_k}L^{(m_k)} + D(P_0||P_1) \leq \epsilon_k\right) \\
&= \mathbb{P}_0\left(D(P_0||P_1) - \frac{1}{m_k}\sum_{j=1}^{m_k}\log\frac{P_0(Y_j)}{P_1(Y_j)} \leq \epsilon_k\right) \\
&\geq 1 - \frac{\sigma^2(P_0||P_1)}{m_k\epsilon_k^2} \qquad (28)
\end{aligned}$$

where the last line follows from Chebyshev's inequality. To insure the probability a null component is discarded on any step is greater than ρ, we have the following condition

$$\epsilon_k \geq \sqrt{\frac{\sigma^2(P_0||P_1)}{m_k(1-\rho)}}$$

As m_k is smallest for $k = 1$, from the definition of m_k in (14)

$$m_k \geq m_1 \geq \frac{m\rho^2 n}{n + sK^2} - 1$$

and the condition is implied for all k provided

$$\epsilon_k \geq \sqrt{\frac{\sigma^2(P_0||P_1)}{\left(\frac{m\rho^2 n}{n+sK^2} - 1\right)(1-\rho)}}. \qquad (29)$$

To summarize developments thus far, we've shown that if $\gamma_k = -m_k(D(P_0||P_1) - \epsilon_k)$ then both

$$\mathbb{P}_1\left(L^{(m_k)} < \gamma_k\right) \leq e^{-m_k(D(P_0||P_1)-\epsilon_k)}$$

and

$$\mathbb{P}_0\left(L^{(m_k)} \leq \gamma_k\right) \geq \rho$$

for any ϵ_k that satisfies (29).

Continuing with m_k as specified in (14),

$$\begin{aligned}
\beta &\leq \sum_{k=1}^{K} \mathbb{P}_1\left(L_{i,k}^{(m_k)} \leq \gamma_k\right) \\
&\leq \sum_{k=1}^{K} e^{-m_k(D(P_0||P_1)-\epsilon_k)} \\
&\leq \sum_{k=1}^{K} \exp\left(-mk \underbrace{\rho^2\left(\frac{n}{n+sK^2}\right)\left(D(P_0||P_1) - \sqrt{\frac{\sigma^2(P_0||P_1)}{\left(\frac{m\rho^2 n}{n+sK^2}-1\right)(1-\rho)}}\right)}_{c_n'}\right) \\
&\leq \frac{e^{-mc_n'}}{1-e^{-mc_n'}} \qquad (30)
\end{aligned}$$

where the last inequality follows as the sum is a geometric series.

With $K = \left\lceil \log_{\frac{1}{1-\rho}}\left(\frac{2(n-s)}{\delta}\right)\right\rceil$, the false positive rate is then

$$\begin{aligned}
\alpha &\leq (1-\rho)^K \\
&\leq (1-\rho)^{\log_{\frac{1}{1-\rho}}\left(\frac{2(n-s)}{\delta}\right)} \\
&\leq \frac{\delta}{2(n-s)} \qquad (31)
\end{aligned}$$

Combining (30) and (31) gives

$$\mathbb{P}_e \leq \frac{\delta}{2} + s\,\frac{e^{-mc_n'}}{1-e^{-mc_n'}}.$$

Next, from the statement of the theorem, let

$$m \geq \frac{\log\left(\frac{4s}{\delta}\right)}{c_n}$$

where

$$c_n = \rho^2\left(\frac{n}{n+sK^2}\right)\left(D(P_0||P_1) - \sqrt{\frac{\sigma^2(P_0||P_1)}{\left(\frac{\rho^2 n \log s}{D(P_0||P_1)(n+sK^2)}-1\right)(1-\rho)}}\right). \qquad (32)$$

Notice $c_n \leq D(P_0||P_1)$. Thus, $m \geq \frac{\log s}{D(P_0||P_1)}$, and

$$m \geq \frac{\log\left(\frac{4s}{\delta}\right)}{c_n} \geq \frac{\log\left(\frac{4s}{\delta}\right)}{c_n'}$$

where

$$c_n' = \rho^2 \left(\frac{n}{n + sK^2} \right) \left(D(P_0||P_1) - \sqrt{\frac{\sigma^2(P_0||P_1)}{\left(\frac{m\rho^2 n}{n+sK^2} - 1 \right)(1-\rho)}} \right).$$

As $m \geq \frac{\log\left(\frac{4s}{\delta}\right)}{c_n'}$,

$$\begin{aligned} \mathbb{P}_e &\leq \frac{\delta}{2} + \frac{\delta/4}{1 - \delta/(4s)} \\ &\leq \delta \end{aligned} \tag{33}$$

where the last inequality holds for $s \geq 1$, $\delta \leq 1$ which proves the theorem. ■