MULTISTAGE ADAPTIVE TESTING OF SPARSE SIGNALS

BY WEINAN WANG, WENGUANG SUN*

University of Southern California

Multistage design has been used in a wide range of scientific fields. By allocating sensing resources adaptively, one can effectively eliminate null locations and localize signals with a smaller study budget. We formulate a decision-theoretic framework for simultaneous multistage adaptive testing and study how to minimize the total number of measurements while meeting pre-specified constraints on both the false positive rate (FPR) and missed discovery rate (MDR). The new procedure, which effectively pools information across individual tests using a simultaneous multistage adaptive ranking and thresholding (SMART) approach, can achieve precise error rates control and lead to great savings in total study costs. Numerical studies confirm the effectiveness of SMART for FPR and MDR control and show that it achieves substantial power gain over existing methods. The SMART procedure is demonstrated through the analysis of high-throughput screening data and spatial imaging data.

1. Introduction. Suppose we wish to recover the support of a \( p \)-vector \( \mu = (\mu_1, \ldots, \mu_p) \in \mathbb{R}^p \) based on measurements from variables \( X_1, \ldots, X_p \). Let \( S = \{i : \mu_i \neq 0\} \) denote the support of \( \mu \). We focus on a setup where measurements on variables are performed sequentially. Consider the following multistage random mixture model [1, 43, 24]:

\[
X_{ij} \sim (1 - \pi)F_0 + \pi F_{1i}; \quad i = 1, \ldots, p,
\]

where the measurements are taken in stages \( j = 1, 2, \ldots, p \), and \( X_{ij} \) follow the null distribution \( F_0 \) if \( i \notin S \) and the non-null distribution \( F_{1i} \) if \( i \in S \). The model assumes that \( F_0 \) is identical for all \( i \notin S \), whereas \( F_{1i} \) can vary across \( i \in S \). Allowing heterogeneous \( F_{1i} \) is desirable in applications where non-zero coordinates have different effect sizes. The mixing proportion \( \pi \) is usually small, and can be understood as follows: \( X_{ij} \) has probability \( 1 - \pi \) of being a null case and probability \( \pi \) of being a signal. Denote \( f_0 \) and \( f_{1i} \) the corresponding densities. The goal of the sparse recovery problem is to find a subset that virtually contains all and only signals.

*The research of W. Sun is supported in part by NSF Grant DMS-CAREER-1255406.

Keywords and phrases: Compound decision problem, distilled sensing, false discovery rate, large-scale multiple testing, missed discovery rate, sequential probability ratio test
1.1. Non-adaptive vs. adaptive designs. Consider a setting where each measurement is associated with a fixed cost; hence collecting many repeated measurements on all variables are prohibitively expensive in large-scale studies where \( p \) is in the order of thousands and even millions.

In applications with a sequential design, it may not be necessary to measure all features at all stages. Let \( \mathcal{A}_j \subset \{1, \ldots, p\} \) denote the set of coordinates where measurements are performed at stage \( j \) and \( X_{\mathcal{A}_j} = (X_{ij} : i \in \mathcal{A}_j) \) the corresponding observations. In a non-adaptive setting, \( X_{ij} \) are sampled at every stage following a pre-fixed policy and each coordinate \( i \) is expected to receive the same amount of measurement budget. In an adaptive sampling design, the sampling scheme varies across different coordinates, with the flexibility of adjusting \( \mathcal{A}_j \) in response to the information collected at previous stages.

The adaptive sampling and inference framework provides a powerful approach to sparse estimation and testing problems. Intuitively, the sensing resources in later stages can be allocated in a more cost-effective way to reflect our updated contextual knowledge during the course of the study; hence greater precision in inference can be achieved with the same study budgets or computational costs. A plethora of powerful multistage testing and estimation procedures have been developed under this flexible framework; some recent developments include the hierarchical testing procedures for pattern recognition [5, 27, 37], distilled sensing and sequential thresholding methods for sparse detection [18, 23, 24], multi-scale search and open-loop feedback control algorithms for adaptive estimation [1, 43, 44] and sequentially designed compressed sensing [19, 25]. These works demonstrate that methodologies adopting adaptive designs can substantially outperform those developed under non-adaptive settings.

Our goal is to develop a cost-effective multistage sampling and inference procedure to narrow down the focus in a sequential manner to identify the vector support reliably. The proposed strategy consists of a stopping rule for selecting the coordinates at which measurements should be performed, together with a testing rule for deciding whether a coordinate contains a signal.

1.2. Applications and statistical challenges. Multistage experiments have been widely used in many scientific fields including environmental sciences [10, 39], microarray, RNA-seq, and protein array experiments [28, 31], geostatistical analysis [30, 7], genome-wide association studies [34, 32] and astronomical surveys [27]. We first describe two applications and discuss important statistical issues in multistage design and analysis.

High-throughput screening (HTS, [46, 6]) is a large-scale hierarchical process that has played a crucial role in fast-advancing fields such as stem cell biology
and drug discovery. In drug discovery, HTS involves testing a large number of chemical compounds in multiple stages including target identification, assay development, primary screening, confirmatory screening, and follow-up of hits. The accurate selection of useful compounds is an important issue at each aforementioned stage of HTS. For example, at the primary screening stage, a library of compounds is tested to generate an initial list of active compounds, or hits. The goal of this stage is to reduce the size of the library significantly with negligible false negative rate. In the confirmatory screening stage, the hits are further investigated to generate a list of confirmed hits, which will be used to generate “leads” for developing drug candidates. As the lab costs for leads generation are very high, the important task at the confirmatory stage is to construct a subset with negligible false positive rate while keeping as many useful compounds as possible. Another application arises from large-scale astronomical surveys for detecting sparse objects of interest. When millions of images are taken with high frequencies, the computational cost of an exhaustive search through every single image and pixel, which often involves testing billions of hypotheses, is prohibitively expensive. A multistage decision process can lead to great savings in total sensing efforts by quickly narrowing down the focus to a much smaller subset of most promising spots in the images [11]. The two applications will be discussed in detail in Section 5.

In the design and analysis of large-scale multistage studies, the inflation of decision errors and soaring study costs are among the top concerns. First, to identify useful signals effectively, we need to control the false negative rate to be small at all stages since missed signals will not be revisited in subsequent stages. Second, to reduce the study costs, it is desirable to eliminate as many locations as possible at each stage. Finally, to avoid misleading scientific conclusions, the final stage of analysis calls for a strict control of the false positive rate. We aim to develop a simultaneous inference framework for multistage decision process to address the above issues integrally.

1.3. Problem formulation. The sequential sparse recovery problem is a compound decision problem [29] where each component problem involves testing a single hypothesis \( H_0 : \mu = 0 \) vs. \( H_1 : \mu \neq 0 \) based on sequentially collected observations. The basic framework is formulated in the seminal work of Wald [41], where the following constrained optimization problem is studied:

\[
(1.2) \quad \text{minimize } E(N) \text{ subject to } \mathbb{P}_{H_0} (\text{Reject}) \leq \alpha' \text{ and } \mathbb{P}_{H_1} (\text{Accept}) \leq \gamma'.
\]

Here \( N \) is the stopping time, and \( \alpha' \) and \( \gamma' \) are pre-specified Type I and Type II error rates. \( E(N) \), which represents the average sampling costs, characterizes the efficiency of a sequential procedure. The sequential probability ratio test
(SPRT) is shown to be optimal \cite{41, 35} for the single sequential testing problem (1.2) in the sense that it has the smallest $\mathbb{E}(N)$ among all sequential procedures at level $(\alpha', \gamma')$.

When many coordinate-wise sequential decisions are made simultaneously, the control of inflated decision errors becomes a critical issue. Denote $\theta_i = \mathbb{I}(\mu_i \neq 0)$, where $\mathbb{I}(\cdot)$ is an indicator function. Let $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)$, where $\delta_i = 0/1$ indicates that $i$ is classified as a null/non-null case. Then the true and estimated supports are denoted $S = (i : \theta_i = 1)$ and $\hat{S} = (i : \delta_i = 1)$, respectively. Define the false positive rate (FPR) and missed discovery rate (MDR) as

\begin{equation}
\text{FPR}(\delta) = \frac{\mathbb{E}\{\sum_{i=1}^{p}(1 - \theta_i)\delta_i\}}{\mathbb{E}(\sum_{i=1}^{p}\delta_i)} \quad \text{and} \quad \text{MDR}(\delta) = \frac{\mathbb{E}\{\sum_{i=1}^{p}\theta_i(1 - \delta_i)\}}{\mathbb{E}(\sum_{i=1}^{p}\theta_i)}.
\end{equation}

The FPR, also referred to as the marginal false discovery rate (mFDR) \cite{14, 36}, is asymptotically equivalent to the widely used false discovery rate (FDR, \cite{3}). The FDR is a powerful and popular error criterion in large-scale testing problems. The main consideration of using FPR (as opposed to FDR) is for analytical convenience. The MDR is also called the “missed rate” \cite{38}. An alternative measure to the MDR is the false negative rate (FNR, \cite{14, 33}). To compare the efficiency, we use the expected average stopping time

\begin{equation}
\text{EAST}(\mathbf{N}) = \mathbb{E}\left(p^{-1}\sum_{i=1}^{p} N_i\right),
\end{equation}

where $\mathbf{N} = (N_1, \cdots, N_p)$, with $N_i$ being the stopping time (or number of measurements) at unit $i$. Let $\alpha$ and $\gamma$ be the pre-specified FPR and MDR levels. We study the following constrained optimization problem:

\begin{equation}
\text{minimize} \ \text{EAST}(\mathbf{N}) \ \text{subject to} \ \text{FPR}(\delta) \leq \alpha \ \text{and} \ \text{MDR}(\delta) \leq \gamma.
\end{equation}

The formulation (1.5) naturally extends the classical formulation (1.2) in \cite{41, 35} to the compound decision setting.

1.4. Main contributions. We formulate the sequential sparse recovery problem as a multiple testing problem with sequentially collected observations. A new adaptive testing procedure is developed under the compound decision-theoretic framework. The proposed procedure, which employs a simultaneous multistage adaptive ranking and thresholding (SMART) approach, not only utilizes all information collected through multiple stages but also exploits the compound decision structure to pool information across different coordinates. We show that SMART controls the FPR and MDR and achieves the information-theoretic lower bounds. Numerical studies confirm the effectiveness of SMART.
for error rates control and demonstrate that it leads to substantial savings in study costs.

SMART has several advantages. First, existing methods such as distilled sensing (DS, [18]) and simple sequential thresholding (SST, [24]) use fixed thresholding rules that do not offer accurate error rate control. Moreover, the literature on sequential multiple testing [2, 16, 45, 17] has focused on the control of false positive findings, and the issues on MDR control and optimal design in the adaptive setting still remain unknown. By contrast, SMART aims to solve the constrained minimization problem (1.5), which addresses the issues on adaptive design, FPR control and MDR control integrally. Second, although existing methods (e.g. [18, 24]) employ a multistage design, the stopping and testing rules at stage $k$ only depend on the observations at the current stage, and the observations from previous stages $j = 1, \cdots, k - 1$ are abandoned. Instead, SMART utilizes all available information from the first stage to the current stage, which leads to more powerful testing and stopping rules. Finally, as opposed to DS and SST that ignore the compound decision structure, SMART employs the ranking and compound thresholding idea in multiple testing to pool information across individual tests. Consequently, SMART controls the error rates more precisely with substantial efficiency gain.

1.5. Organization of the paper. In Section 2 we first formulate a compound decision-theoretic framework for sequential sparse recovery problems, and then propose oracle and SMART rules for FPR and MDR control. Section 3 derives the fundamental limits for sparse inference and establishes the asymptotic optimality of the proposed SMART procedure. In Section 4, we conduct numerical comparisons with competitive methods to demonstrate the merits of SMART. Section 5 applies SMART for analyzing data from high-throughput screening studies. The theorems are proved in Section 6. Additional analytical and numerical results are provided in the Supplementary Material.

2. Oracle and SMART Rules for Sparse Recovery. The sparse recovery problem in the adaptive setting involves the simultaneous testing of $p$ hypotheses:

\begin{equation}
H_{0,i} : \mu_i = 0 \text{ vs. } H_{1,i} : \mu_i \neq 0, \quad 1 \leq i \leq p,
\end{equation}

based on $p$ streams of observations. In this section, we first formulate the sequential sparse recovery problem in a decision-theoretic framework (Section 2.1), then derive an oracle procedure for FPR and MDR control (Section 2.2), and finally propose new methodologies to approximate the oracle procedure (Sections 2.3-2.4).
2.1. A decision-theoretic formulation. Let $X_{ij}$ be the measurement of variable $X_i$ at stage $j$ and $X_i^j = (X_{i1}, \ldots, X_{ij})$ the collection of measurements on $X_i$ up to stage $j$. A multistage decision procedure involves choosing a stopping rule and a testing rule for each location. At location $i$, the stopping rule $\tau_i$ consists of a series of functions $\tau_{i1}(X_i^1), \tau_{i2}(X_i^2), \ldots$, where $\tau_{ij}$ takes values in $\{0, 1\}$, with 0 and 1 standing for “taking another observation” and “stopping sampling and making a decision”, respectively. We consider a class of multistage designs where the focus is sequentially narrowed down, i.e. the active sets satisfy $A_1 = \{1, \ldots, p\}$, and $A_j \subset A_{j-1}$ for $j = 2, 3, \ldots$. It follows that at every coordinate $i \in A_j$, there are three possible actions: (i) stop sampling and claim $H_{0,i}$ is true; (ii) stop sampling and claim $H_{0,i}$ is false; and (iii) do not make a decision and take another observation. Define the stopping time $N_i = \min\{n \geq 1 : \tau_{in}(X_i^n) = 1\}$. Then the stopping rule $\tau_i$ can be equivalently described by a stopping time $N_i$. The testing rule $\delta_i \in \{0, 1\}$ is carried out at stage $N_i$ (the terminal sampling stage), where $\delta_i = 0/1$ indicates that $i$ is classified as a null/non-null case. A multistage decision procedure is therefore denoted $d = (N, \delta)$, where $N = (N_1, \ldots, N_p)$ and $\delta = (\delta_1, \ldots, \delta_p)$ are the stopping times and terminal decisions, respectively.

Assume that the cost of taking one observation is $c$. From a decision-theoretic viewpoint, we can study a weighted classification problem with the following loss function:

$$L_{\lambda_1, \lambda_2}(\theta, d) = \lambda_1 \{\sum_{i=1}^p (1 - \theta_i)\delta_i\} + \lambda_2 \{\sum_{i=1}^p \theta_i(1 - \delta_i)\} + c \sum_{i=1}^p N_i,$$

where $\lambda_1$ and $\lambda_2$ are the costs for a false positive decision and a false negative decision, respectively. The sum of the first two terms in (2.2) corresponds to the total decision errors, and the last term gives the total sampling costs. The optimal solution to the weighted classification problem is the Bayes sequential procedure $d^\ast$ that minimizes the expected loss

$$E\{L_{\lambda_1, \lambda_2}(\theta, d^\ast)\} = \inf_d E\{L_{\lambda_1, \lambda_2}(\theta, d)\}.$$

In [4], $d^\ast$ is represented by a thresholding rule based on the oracle statistic

$$T_{\delta_i}^{i,j} = \mathbb{P}(\theta_i = 0 | X_i^j).$$

$T_{\delta_i}^{i,j}$ is the posterior probability of case $i$ being a null given $X_i^j$. We view $T_{\delta_i}^{i,j}$ as a significance index reflecting our confidence on claiming $H_{0,i}$ is true based on all information available at stage $j$.

In practice we only specify the FPR and MDR levels $\alpha$ and $\gamma$ but do not know $\lambda_1$ and $\lambda_2$. However, the optimal solution to the weighted classification problem (2.3) motivates us to conjecture that the optimal solution to the sparse recovery problem (1.5) should also be a thresholding rule based on $T_{\delta_i}^{i,j}$. This conjecture will be established rigorously in the next subsection.
2.2. Oracle procedure. Let $t_l$ and $t_u$ be constants satisfying $0 \leq t_l < t_u \leq 1$. Consider a class of sequential testing procedures $d^x(t_l, t_u)$ of the form:

\[
\text{stop sampling for unit } i \text{ at } N_i = \min\{j \geq 1 : T_{i,j} \leq t_l \text{ or } T_{i,j} \geq t_u\},
\]

(2.5) and decide $\delta_i = 1$ if $T_{i,j} \leq t_l$ and $\delta_i = 0$ if $T_{i,j} \geq t_u$.

Denote by $D_{\alpha, \gamma}$ the collection of all sequential decision procedures which simultaneously satisfy (i) $\text{FPR}(d) \leq \alpha$ and (ii) $\text{MDR}(d) \leq \gamma$. The following assumption is a standard condition in the sequential analysis literature (e.g. [4, 35]). It essentially requires that $f(X_i|\theta_i=0)$ and $f(X_i|\theta_i=1)$ differ with some positive probability, which ensures that the sequential testing procedure has a finite stopping time.

Assumption 1. Suppose the dimension $p$ is fixed. Assume that $\mathbb{P}_{\theta_i}(Z_{i,1} = 0) < 1$ for $i = 1, \cdots, p$, where $Z_{i,1} = \log \left( \frac{f(X_i|\theta_i=1)}{f(X_i|\theta_i=0)} \right)$.

The next theorem derives an oracle sequential testing procedure that provides the optimal solution to the constrained optimization problem (1.5).

**Theorem 1.** Consider a class of sequential testing rules $d^x(t_l, t_u)$ taking the form of (2.5). Denote $Q_{OR}(t_l, t_u)$ and $\tilde{Q}_{OR}(t_l, t_u)$ the FPR and MDR levels of $d^x(t_l, t_u)$, respectively. Then under Assumption 1, we have

(a). $Q_{OR}(t_l, t_u)$ is non-decreasing in $t_l$ for a fixed $t_u$, and $\tilde{Q}_{OR}(t_l, t_u)$ is non-increasing in $t_u$ for a fixed $t_l$.

(b). Let $0 < \alpha, \gamma < 1$. Then there exists a pair of oracle thresholds $(t_l^{OR}, t_u^{OR})$, based on which we can define the oracle procedure

\[
d_{OR} \equiv d^x(t_l^{OR}, t_u^{OR}),
\]

(2.6) such that $d_{OR}$ is optimal in the sense that (i) $\text{FPR}(d_{OR}) \leq \alpha$; (ii) $\text{MDR}(d_{OR}) \leq \gamma$; and (iii) $\text{ESS}(d_{OR}) \leq \text{ESS}(d_*)$ for all $d_* \in D_{\alpha, \gamma}$.

The oracle procedure $d_{OR}$ is a thresholding rule based on the oracle statistic $T_{OR}$ and oracle thresholds $t_l^{OR}$ and $t_u^{OR}$. However, Theorem 1 only shows the existence of $t_l^{OR}$ and $t_u^{OR}$, which are unknown in practice. In Section 2.3, we extend the classical ideas in [41, 35] to derive approximations of $t_l^{OR}$ and $t_u^{OR}$. The approximations guarantee conservative FPR and MDR control but suffer from the “overshoot” problem in sequential testing. Section 2.4 develops a ranking and compound thresholding algorithm to improve the approximation.
2.3. Approximation of oracle thresholds and its properties. If we focus on a single location \( i \), then the sequential probability ratio test (SPRT) \([41]\) is a thresholding rule based on 
\[
L_{i,j} = \frac{f(X_j|\theta_i=1)}{f(X_j|\theta_i=0)}
\]
and of the form

if \( \log L_{i,j} \leq a \), stop sampling and decide \( H_{0,i} \) is true;

if \( \log L_{i,j} \geq b \), stop sampling and decide \( H_{0,i} \) is false;

if \( a < \log L_{i,j} < b \), take another observation.

Let \( \alpha' = \mathbb{P}_{H_{0,i}}(\text{Reject } H_{0,i}) \) and \( \gamma' = \mathbb{P}_{H_{1,i}}(\text{Accept } H_{0,i}) \) be prespecified Type I and Type II error rates. Then by applying Wald’s identify \([42, 35]\), the thresholds \( a \) and \( b \) can be approximated as:

\[
(2.7) \quad \tilde{a} = \log \frac{\gamma'}{1 - \alpha'} \quad \text{and} \quad \tilde{b} = \log \frac{1 - \gamma'}{\alpha'}.
\]

The oracle statistic \( T_{OR}^{i,j} = \frac{(1 - \pi)}{(1 - \pi) + \pi L_{i,j}} \) is monotone in \( L_{i,j} \). Therefore in the multiple hypothesis setting, we can view the oracle rule \( d_{OR} = d^*(\tilde{t}_{OR}^l, \tilde{t}_{OR}^u) \) as \( m \) parallel SPRTs. Therefore the problem boils down to how to obtain approximate formulas of the oracle thresholds \( \tilde{t}_{OR}^l \) and \( \tilde{t}_{OR}^u \) for a given pair of pre-specified FPR and MDR levels \( (\alpha, \gamma) \). In our derivation, the classical techniques (e.g. Section 7.5.2 in \([4]\)) are used, and the relationships between the FPR and MDR levels \( (\alpha, \gamma) \) and the Type I and Type II error rates \( (\alpha', \gamma') \) are exploited. The derivation of approximation formulas with the FPR and MDR constraints is complicated; we provide detailed arguments in Appendix A. The approximated thresholds are given by

\[
(2.8) \quad \tilde{t}_{OR}^l = \alpha \quad \text{and} \quad \tilde{t}_{OR}^u = \frac{1 - \pi}{\pi \gamma + 1 - \pi}.
\]

The next theorem shows that the pair \((2.8)\) is valid for FPR and MDR control.

**Theorem 2.** Consider multistage model (1.1). Denote \( \tilde{d}_{OR} = d^*(\tilde{t}_{OR}^l, \tilde{t}_{OR}^u) \) the thresholding procedure that operates according to (2.5) with upper and lower thresholds given by (2.8). Then \( \text{FPR}(\tilde{d}_{OR}) \leq \alpha \) and \( \text{MDR}(\tilde{d}_{OR}) \leq \gamma \).

The approximations in (2.8) lead to very conservative FPR and MDR levels. The main issue is the “overshoot” problem, i.e. in the parallel SPRTs the boundaries are not hit exactly. The conservativeness in the error rates control would result in increased study budgets. The situation is much exacerbated in high-dimensional settings where most locations are only tested once or twice. In fact, the “step sizes” of many SPRTs tend to be quite large relative to the given boundaries and the overshoot problems would occur with high frequencies. Our
simulation studies reveal that the actual FPR and MDR levels are often only half of the nominal levels, resulting in substantial efficiency loss. The above concern motivates us to develop more powerful methodologies to improve the approximations in (2.8).

2.4. The SMART procedure. The oracle procedure $d_{OR} (2.6)$ can be interpreted in two ways: one is a collection of $p$ parallel and independent SPRTs, and the other is a stage-wise simultaneous inference procedure. This section takes the latter view. To gain more insights on the overall structure of the problem, we rewrite (2.2) as the sum of stage-wise losses:

\[
L_{\lambda_1, \lambda_2}(\theta, d) = \sum_{j=1}^{N} \left[ \sum_{i \in S_j} \{ \lambda_1 (1 - \theta_i) \delta_i + \lambda_2 \theta_i (1 - \delta_i) \} + c \cdot \text{Card} (A_j) \right],
\]

where $N = \max \{ N_i : 1 \leq i \leq p \}$ and $S_j = \{ i : N_i = j \}$. At stage $j$, we stop sampling on $S_j \subset A_j$ and make terminal decisions at every $i \in S_j$; the remaining locations will become the active set for the next stage $A_{j+1} = A_j \setminus S_j$, on which new observations are taken. We then proceed to make further decisions on $A_{j+1}$. The process will be repeated until the active set becomes empty.

This simultaneous decision view motivates us to employ the idea of ranking and thresholding that has been widely used in the multiple testing literature. For example, the FDR procedure in [3] first orders all $p$-values from the smallest to largest, and then uses a step-up method to choose a cutoff along the $p$-value ranking. Under (2.9), we make three types of decisions simultaneously at each stage based on the ranking produced by the ordered $T_{OR}$: (i) identifying non-null cases, (ii) eliminating null cases, and (iii) selecting coordinates for further measurements. Thus the proposed multistage testing procedure similarly operates via a ranking and thresholding scheme. We describe in Table 1 the proposed algorithm, called SMART, for “simultaneous multistage adaptive ranking and thresholding.”

The operation of SMART can be described as follows. At stage $j$, we first calculate the oracle statistics in the active set $A_j$ and sort them from smallest to largest. Then we carry out two thresholding procedures along the ranking: algorithm (a) chooses a lower cutoff $t_{OR}^{l,j}$ with selected locations claimed as signals; algorithm (b) chooses an upper cutoff $t_{OR}^{u,j}$ with selected locations claimed as nulls. We stop sampling on locations where definitive decisions are made, and take new observations on remaining locations for further investigation. The above steps will be iterated until convergence.

The SMART algorithm utilizes stage-wise thresholds $t_{OR}^{l,j}$ and $t_{OR}^{u,j}$. The operation of the algorithm implies that we always have $t_{OR}^{l,j} \leq t_{OR}^{l}$ and $t_{OR}^{u,j} \geq t_{OR}^{l}$.
Algorithm 1. The SMART procedure

Define the lower and upper thresholds \( \tilde{t}_\text{OR} = \alpha \) and \( \tilde{t}_\text{OR} = \frac{1-\pi}{\pi+1-\pi} \).

Let \( \mathcal{A}_j \) be the active set at stage \( j \), \( j = 1, 2, \cdots \).

Iterate Step 1 to Step 3 until \( \mathcal{A}_j = \emptyset \).

**Step 1 (Ranking).** For all \( i \in \mathcal{A}_j \), compute \( T_{ij} \) and sort them in ascending order \( T_{(1),j} \leq T_{(2),j} \leq \cdots \leq T_{(k_j),j} \), where \( k_j = \text{Card}(\mathcal{A}_j) \).

**Step 2 (Thresholding).**

(a) (Signal discovery). Let \( k_{s,j} = \max \{ r : r^{-1} \sum_{i=1}^{r} T_{(i),j} \leq \tilde{t}_{\text{OR}} \} \) and \( \hat{t}_{\text{OR}} = T_{(k_{s,j}),j} \).

Define \( S_{s,j} = \{ i \in \mathcal{A}_j : T_{ij} \leq \hat{t}_{\text{OR}} \} \). For all \( i \in S_{s,j} \), stop sampling and let \( \delta_i = 1 \).

(b) (Noise elimination). Let \( k_{e,j} = \max \{ r : r^{-1} \sum_{i=0}^{r-1} (T_{(k_j-i),j} - T_{(i),j}) \geq \tilde{t}_{\text{OR}} \} \) and \( \hat{t}_{\text{OR}} = T_{(k_j-k_{e,j}+1),j} \).

Define \( S_{e,j} = \{ i \in \mathcal{A}_j : T_{ij} \geq \hat{t}_{\text{OR}} \} \).

For all \( i \in S_{e,j} \), stop sampling and let \( \delta_i = 0 \).

**Step 3 (Updating).** Let \( \mathcal{A}_{j+1} = \mathcal{A}_j \setminus (S_{s,j} \cup S_{e,j}) \). Take new observations on \( \mathcal{A}_{j+1} \).

for all \( j \). Hence SMART always uses fewer samples than \( \tilde{d}_{\text{OR}} \). The next theorem shows that SMART is valid for FPR and MDR control.

**THEOREM 3.** Denote \( d_{SM} \) the SMART procedure described in Algorithm 1 with pre-specified FPR and MDR levels \( (\alpha, \gamma) \). Then

\[
\text{FPR}(d_{SM}) \leq \alpha \quad \text{and} \quad \text{MDR}(d_{SM}) \leq \gamma.
\]

In the thresholding step, SMART chooses the lower and upper cutoffs based on the moving averages of the selected oracle statistics. We call SMART a compound thresholding procedure, for its stage-wise cutoffs \( \hat{t}_{\text{OR}}^{(j)} \) and \( \hat{t}_{\text{OR}}^{(j)} \) are jointly determined by data from multiple locations. By contrast, we call \( \tilde{d}_{\text{OR}} \) a simple thresholding rule as the decision at location \( i \) only depends on its own data. By adopting a compound thresholding scheme and pooling information across parallel SPRTs, SMART overcomes the overshoot problem of individual SPRTs. To illustrate the point, consider the following toy example.

**EXAMPLE 2.1.** Let the FDR level \( \alpha = 0.05 \). Suppose at stage \( j \) the ordered \( T_{OR} \) values are \( \{0.01, 0.055, 0.07, 0.10, \cdots \} \). If we use the simple thresholding rule \( \tilde{d}_{OR} \) with \( \tilde{t}_{OR} = \alpha = 0.05 \), then we can reject one hypothesis with \( T_{(1),j} = 0.01 \); the gap between \( T_{(1),j} \) and \( \alpha \) is 0.04. By contrast, the moving average of the top three statistics is 0.045 < \( \alpha \); hence SMART rejects three hypotheses. The gap between the moving average and the threshold is only 0.005. Thus the
boundary $\alpha$ can be hit more precisely by the moving average. Therefore SMART has smaller approximation errors compared to individual SPRTs, which leads to more accurate FPR and MDR control as well as savings in study budgets.

2.5. Connections to existing ideas. In the sparse recovery problem (1.5), the $p$ sequential decisions are combined and evaluated as a whole; this is called a compound decision problem [29]. Let $d = (d_1, \ldots, d_p)$ denote a decision procedure. Then $d$ is simple if $d_i$ depends on data from location $i$ alone, and compound if $d_i$ also depends on data from other locations $j \neq i$. A fundamental result in [29] is that compound rules are in general superior to simple rules, even when the observations from different units are independent. In the context of multiple testing, Sun and Cai (2007) [36] showed that more powerful FDR procedures can be constructed by pooling information from independent tests. The proposed SMART procedure further reveals that the accuracy of single SPRTs can be greatly improved by exploiting the overall structure of a compound decision problem.

Compound thresholding is a popular and powerful idea in multiple testing [3, 36]. For example, the Bonferroni method can be uniformly improved by compound thresholding rules such as Holm’s procedure [21] and Hochberg’s procedure [20]. Moreover, the well known BH procedure [3] is also a compound thresholding procedure. To see this, let $G(t)$ be the empirical distribution of $p$-values. Then the BH threshold, given by $t_{BH}(\hat{G}) = \sup\{t : t/\hat{G}(t) \leq \alpha\}$ [14], depends on a bulk of $p$-values. The aforementioned methods employ step-wise algorithms that involve ranking and compound thresholding, which is also adopted by SMART. By pooling information between different locations, the approximation errors of simple thresholding rules can be greatly reduced.

The distilled sensing (DS, [18]) and single sequential thresholding (SST, [24]) methods provide efficient adaptive sampling schemes that narrow down the focus in a sequential manner. SMART employs a similar adaptive sampling and inference framework but improves DS and SST in several ways. First, as opposed to DS and SST that use fixed thresholding rules at all units and through all stages [e.g. DS always uses 0 as the threshold and hence eliminates roughly half of the coordinates in each distillation stage], SMART uses data-driven thresholds that are adaptive to both data and pre-specified FPR and MDR levels. Second, the sampling and inference rules in DS and SST only utilize observations at the current stage. By contrast, the building block of SMART is $T_{OR}^{i,j}$, which utilizes all observations collected up to the current stage. This can greatly increase the signal to noise ratio and lead to more powerful testing and stopping rules. Third, SMART utilizes compound thresholding to exploit the multiple hypothesis structure, which can greatly improve the accuracy of the
error rates control. Finally, DS and SST only have a stopping rule to eliminate null locations. Intuitively, such a scheme is inefficient as one should also stop sampling at locations where the evidence against the null is extremely strong. The proposed scheme in Step 2 of SMART involves two algorithms that can simultaneously identify signals and eliminate noise. In other words, SMART stops sampling once a definitive decision ($\delta_i = 0$ or $\delta_i = 1$) is reached; this more flexible operation is desirable and would further save study budgets.

3. Limits of Sparse Recovery. This section discusses the notion of fundamental limits in sparse inference for both fixed and adaptive settings. The discussion serves two purposes. First, as the limits reflect the difficulties of support recovery under different settings, they demonstrate the advantage of using adaptive designs over fixed designs. Second, the limit provides an optimality benchmark for what can be achieved asymptotically by any inference procedure in the form of a lower bound; hence we can establish the optimality of an inference procedure by proving its capability of achieving the limit.

3.1. The fundamental limits in adaptive designs. Different from previous sections that consider a finite (and fixed) dimension $p$, the asymptotic analysis in this section assumes that $p \to \infty$. Under this asymptotic framework, the FPR and MDR levels are denoted $\alpha_p$ and $\gamma_p$, both of which tend to zero as $p \to \infty$. In Malloy and Nowark (2014) [24], the fundamental limits for reliable recovery were established under the family wise error rate (FWER) criterion. This section extends their theory to the important FPR and MDR paradigm.

The basic setup assumes that the null and alternative distributions $F_0$ and $F_1$ are identical across all locations. This more restrictive model is only for theoretical analysis and our proposed SMART procedure works under the more general model (1.1), which allows $F_{1i}$ to vary across testing units. Let $G$ and $H$ be two distributions with corresponding densities $g$ and $h$. Define $D(G\|H) = \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{h(x)} \, dx$. Then the Kullback-Leibler (KL) divergence, given by

$$D_{KL}(F_0, F_1) = \max\{D(F_0\|F_1), D(F_1\|F_0)\},$$

can be used to measure the distance between $F_0$ and $F_1$. Let $\tau$ be the average number of measurements allocated to each unit. Consider a general multistage decision procedure $d$. The performance of $d$ is characterized by its total risk $R^*(d) = \text{FPR}(d) + \text{MDR}(d)$. Intuitively, $D_{KL}(F_0, F_1)$ and $\tau$ together would characterize the possibility of constructing a $d$ such that $R^*(d) \to 0$.

A decision rule $d$ is symmetric if $d(\Psi(x_i^j : i \in A_j)) = \Psi\{d(x_i^j : i \in A_j)\}$ for all permutation operators $\Psi$ ([9]). Most existing multistage testing methods, such as DS, SST and SMART, are symmetric procedures. The fundamental
limit, described in the next theorem, gives the minimum condition under which it is possible to construct a symmetric $d$ such that $R^*(d) \to 0$.

**Theorem 4. Fundamental limits (lower bound).** Let $d$ be a symmetric multistage adaptive testing rule. Assume that $\pi < \frac{1}{3}$. If $\tau \equiv \text{EAST}(d) \leq \frac{1}{D_{KL}(F_0, F_1)} \log(4\eta)^{-1}$, then we must have $R^*(d) \geq \eta$ for all $\eta > 0$.

In asymptotic analyses we typically take an $\eta$ that converges to 0 slowly. Theorem 4 shows that any adaptive procedure with total risk tending to zero must at least have an EAST (or the average sample size per dimension) in the order of $\log\{4\eta\}^{-1}$. This limit can be used as a theoretical measure to assess the efficiency of a multistage procedure. Denote $d_{SM} = (N_{SM}, \delta_{SM})$ the SMART procedure described in Algorithm 1. As we proceed, we need the following assumption that is essentially equivalent to Condition (8) in [24]:

\begin{equation}
(3.1) \quad E(T_{i,N_i}^{i,N_i} | T_{i,N_i}^{i,N_i} > t_u) \leq \frac{t_u}{(1 - t_u)e^{-C_1} + t_u}, \quad E(T_{i,N_i}^{i,N_i} | T_{i,N_i}^{i,N_i} < t_l) \geq \frac{t_l}{(1 - t_l)e^{C_2} + t_l}
\end{equation}

for all possible thresholds $t_l, t_u$. The condition is satisfied when $\log \mathcal{L}_i,1$ follows a bounded distribution such Gaussian and exponential distributions. A more detailed discussion on this issue can be found in [15]. The following theorem derives the upper bound and establishes the optimality of SMART.

**Theorem 5. Asymptotic optimality (Upper bound).** Consider the SMART procedure described in Algorithm 1 with lower and upper thresholds

\begin{equation}
t_l = \frac{1}{f(p)^{1+\epsilon}} \quad \text{and} \quad t_u = \frac{(1 - \pi)f(p)^{1+\epsilon}}{\pi + (1 - \pi)f(p)^{1+\epsilon}},
\end{equation}

where $\epsilon > 0$ is a small constant and $f(p)$ is a function of $p$ that grows to infinity at an arbitrarily slow rate. Then under (3.1), the SMART procedure satisfies $\lim_{p \to \infty} R^*(\delta_{SM}) = 0$ and

\begin{equation}
\lim_{p \to \infty} \text{EAST}(N_{SM}) \leq \frac{(1 + \epsilon)\log f(p)}{\min\{D(F_0|F_1), D(F_1|F_0)\}}.
\end{equation}

If we take $\eta = f(p)$ with $f(p) \to \infty$ slowly, then the rates in Theorems 4 and 5 would match. Therefore SMART is optimal for adaptive testing under this asymptotic regime.

3.2. Comparison with existing results. We first review the limits for fixed and adaptive designs in the literature and then compare them with our new limits. Consider a two-point normal mixture model under a single-stage design

\begin{equation}
X_i \sim^d (1 - \pi_p)N(0,1) + \pi_pN(\mu_p, \sigma^2),
\end{equation}
where \( \pi_p = p^{-\beta} \), \( \mu_p = \sqrt{2r \log p} \) and \( 0 < \beta < 1 \), \( 0 < r < 1 \). The fundamental limits for a range of global and simultaneous inference problems have derived under this setup [12, 8, 40]. Of particular interest is the classification boundary [26, 18, 40], which demarcates the possibility of constructing a subset with both the FPR and MDR tending to zero. Under model (3.2), the classification boundary is a straight line \( r = \beta \) in the \( \beta-r \) plane for both the homoscedastic case (\( \sigma = 1 \)) [26, 18] and heteroscedastic case (\( \sigma \neq 1 \)) [40]. Hence the goal of \( R^* \to 0 \) requires that the signal magnitude \( \mu_p \) must be at least in the order of \( \sqrt{\log p} \). This gives the fundamental limit of sparse recovery for fixed designs.

The rate \( \sqrt{\log p} \) can be substantially improved in the adaptive setting. For example, Haupt et al. (2011) [18] proposed the distilled sensing (DS) method, which is capable of achieving the classification boundary with much weaker signals. DS is a multistage testing procedure with a total measurement budget of \( 2p \). It assumes that observations follow a mixture model with noise distributed as standard normal. At each stage, DS keeps locations with positive observations and obtain new observations for these locations in the next stage. It was shown in [18] that after \( k = \max\{\lceil \log_2 \log p \rceil, 0\} + 2 \) steps, the DS algorithm successfully constructs a subset with both FPR and MDR tending to zero provided that \( \mu_p \) diverges at an arbitrary rate in the problem dimension \( p \). A more general result on the limit of sparse recovery was obtained in [24], where the KL divergence and the average measurements per dimension \( \tau \) are used in place of the growing signal amplitude \( \mu_p \) to characterize the difficulty of the problem. Let \( s \) denote the cardinality of the support. It was shown in [24] that under fixed designs, the reliable recovery requires that \( \tau \) must be at least in the order of \( \log p \), whereas under adaptive designs, the required \( \tau \) is in the order of \( \log s \). This reveals the advantage of adaptive designs (note \( \log s \sim \log \log p \) under the calibration \( \pi_p = p^{-\beta} \)). SST is optimal in the sense that it achieves the rate of \( \log s \) asymptotically [24].

The upper and lower bounds presented in Theorems 4 and 5 show that the FPR-MDR paradigm requires fewer samples to guarantee that \( R^* \to 0 \). The result is consistent with the rate achieved by DS under model (3.2) [18]. SMART only requires a \( \tau \) of the order \( \log f(p) \), where \( f(p) \to \infty \) at any rate. This slightly improves the rate of \( \log s \) achieved by SST. The improvement is expected as SST is developed to control the more stringent FWER.

4. Simulation. Section 4.1 discusses the estimation of the oracle statistic \( T_{\text{OR}} \) under a hierarchical normal mixture model. Sections 4.2 and 4.3 compare SMART with competitive methods.

4.1. Estimation of the oracle statistic. In practice the oracle statistic \( T_{\text{OR}} \) is unknown and needs to be estimated from data. We present detailed formulæ
for computing $T_{OR}$ in a Bayesian hierarchical model considered in [1, 43]. The model, which utilizes non-informative priors and allows varied signal magnitudes across locations, has been widely used in signal processing by providing a flexible framework for a range of estimation and testing problems.

Let $\theta_1, \ldots, \theta_p$ be independent Bernoulli($\pi$) variables. Assume that observations $X_{ij}$ obey the following multistage model:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

where $\mu_i = 0$ if $\theta_i = 0$, and $\mu_i \sim N(\eta_i, \tau_i^2)$ if $\theta_i = 1$. The prior mean and variance are denoted $\eta_i(0)$ and $\tau_i^2(0)$, respectively. We use the method in Jin and Cai (2007) [22] to get an initial estimate of the prior probability $\hat{\pi}$ and the null distribution parameter $\hat{\sigma}^2$. Our initialization utilizes non-informative priors: for all $i$ let $T_{OR}^{i,0} = \hat{\pi}$, $\eta_i(0)$ be the average of top 100$\hat{\pi}$% observations, and $\tau_i^2(0) = 1$.

At each stage $j$, we collect new observations on the active set and update our estimates of $(\pi_i, \eta_i, \tau_i^2)$ by $\{\pi_i(j), \eta_i(j), \tau_i^2(j)\}$. Let $\phi(\cdot; \mu, \sigma^2)$ denote the density function of a normal distribution with mean $\mu$ and variance $\sigma^2$. The recursive formula is based on $f(\mu_i|\theta_i = 1, x_i^{j+1}) \propto f(x_i^{j+1}|\mu_i)f(\mu_i|x_i^j, \theta_i = 1) = \phi(x_i^{j+1}; \mu_i, \sigma^2)\phi(\mu_i; \eta_i(j), \tau_i^2(j))$. Completing the squares, we have $f(\mu_i|\theta_i = 1, x_i^{j+1}) \propto \exp \left[-\frac{(x_i^{j+1} - \mu_i - \eta_i(j+1))^2}{2\tau_i^2(j+1)}\right]$. The proposed algorithm operates as follows.

**Algorithm 2. Recursive estimation of $T_{OR}^{i,j}$**

For $j = 0, 1, 2, \ldots, N$, compute the following quantities in turn:

$$\eta_i(j + 1) = \frac{\tau_i^2(j)}{\tau_i^2(j) + \hat{\sigma}^2} x_{i,j+1} + \frac{\hat{\sigma}^2}{\tau_i^2(j) + \hat{\sigma}^2} \eta_i(j);$$

$$\tau_i^2(j + 1) = \frac{\tau_i^2(j) \hat{\sigma}^2}{\tau_i^2(j) + \hat{\sigma}^2};$$

$$f_{i,0,j+1} = \phi(x_i^{j+1}; 0, \hat{\sigma}^2), \quad f_{i,1,j+1} = \phi(x_i^{j+1}; \eta_i(j), \tau_i^2(j) + \hat{\sigma}^2);$$

$$T_{OR}^{i,j+1} = \mathbb{P}(\theta_i = 0|x_i^{j+1}) = \frac{T_{OR}^{i,j} f_{i,0,j+1}}{T_{OR}^{i,j} f_{i,0,j+1} + (1 - T_{OR}^{i,j}) f_{i,1,j+1}}.$$  

4.2. Simulation 1: simple thresholding vs. compound thresholding. This simulation study compares the following methods: (i) the single thresholding procedure $\delta_{ST}$ that assumes an oracle knows the true parameters (OR.ST); (ii) the SMART procedure $\delta_{SM}$ with known parameters (OR.SM); (iii) $\delta_{OR}$ with
parameters estimated via Algorithm 2 (DD.ST), where DD refers to “data-driven”; (iv) $\delta_{SM}$ with estimated parameters (DD.SM). The main purpose is to show the advantage of compound thresholding. Specifically, $\hat{\delta}_{OR}$ operates as $p$ parallel SPRTs and suffers from the overshoot problem. The approximation errors of SPRTs can be greatly reduced by SMART, which control the FPR and MDR more accurately with smaller sample sizes.

We generate data from the multistage model (4.1). The number of locations is $p = 10^5$. Let $(\alpha, \gamma) = (0.05, 0.05)$ be pre-specified FPR and MDR levels. The following settings are considered:

Setting 1: $\pi = 0.01$. $\eta_i = \mu_0$ for all $i$. Vary $\mu_0$ from 2 to 4 with step size 0.2.
Setting 2: $\pi = 0.05$. $\eta_i = \mu_0$ for all $i$. Vary $\mu_0$ from 2 to 4 with step size 0.2.
Setting 3: Draw $\eta_i$ randomly from a uniform distribution $U(2, 4)$. Vary $\pi$ from 0.05 to 0.2 with step size 0.01.

We apply the four methods to the simulated data. The FPR, MDR and ESS (expected sample size) are computed based on the average of 100 replications, and are plotted as functions of varied parameter values. The results are summarized in Figure 1. In Setting 3, the two oracle methods (OR.ST and OR.SM) are not implemented as recursive formulae for $T_{OR}$ are unavailable.

We can see that all four methods control the FPR at the nominal level. However, the two single thresholding methods (OR.TH and DD.TH) are very conservative (the actual FPR is only about half of the nominal level). Similarly, both OR.TH and DD.TH are very conservative for MDR control. In contrast, the SMART procedures (OR.SM and DD.SM) control the error rates more accurately and require smaller sample sizes. When signals are sparse and weak (top middle panel), the MDR level of DD.SM is slightly higher than the nominal level. This is due to the estimation errors occurred at stage 1. It is of interest to develop more accurate estimation procedures in such settings.

4.3. Simulation 2: SMART vs. distilled sensing. This simulation study compares SMART and DS [18]. As DS does not provide precise error rates control, the simulation is designed in the following way to make the comparison on an equal footing: we first run DS with up to 10 stages and record its FPR and MDR levels. Then we apply SMART at the corresponding FPR and MDR levels so that the two methods have roughly equal error rates. The ESS is used to compare the efficiency.

The data are generated from the multistage model (4.1). The number of locations is $p = 10^5$. The following two settings are considered:

Setting 1: $\pi = 0.05$. $\eta_i = \mu_0$ for all $i$. Vary $\mu_0$ from 2 to 4 with step size 0.2.
Setting 2: Draw $\eta_i$ randomly from $U(2, 4)$. Vary $\pi$ from 0.05 to 0.2.
In both settings, the FPR, MDR and ESS are computed by averaging over 100 replications, and are plotted as functions of varied parameter values. The results are summarized in Figure 2.

We can see that the error rates of both OR.SM and DD.SM match well with those of DS, but they require fewer samples. OR.ST and DD.ST also outperform DS, achieving lower error rates with fewer samples.

Fig 1: Comparison with single thresholding. The displayed procedures are DD.SM (●), OR.SM (■), DD.ST (▲), OR.ST (+).

5. Applications. This section applies SMART and DS to the HTS studies. The other application to satellite image analysis is provided in the Supplementary Material. We compare the performances of different methods in two ways: (i) the total sample sizes needed to achieve pre-specified error rates, and (ii) the actual error rates achieved for a fixed total sample size.
Fig 2: Comparison with DS. The displayed procedures are DS (●), DD.SM (▲), OR.SM (+), DD.ST (■), OR.ST (⊗).

The goal of the HTS study conducted by McKoy et al. (2012) is to identify novel inhibitors of the amyloid beta peptide (Aβ), whose aggregation is believed to be a trigger of the Alzheimer’s disease. In the study, a total of $p = 51,840$ compounds are tested, with three measurements recorded for each compound. We use the observed data set as a pilot data set and simulate observations in later stages to illustrate how to design a multistage sampling and inference procedure for identifying useful compounds.

We first obtain z-scores based on the average of the three measurements and then estimate the non-null proportion and null distribution using the method in [22]. The estimated non-null proportion is $\hat{\pi} \approx 0.0007$, and the estimated null distribution (referred to as the empirical null distribution, [13]) is $N(\hat{\mu}_0, \hat{\sigma}_0^2)$ with $\hat{\mu}_0 = 0.2459, \hat{\sigma}_0 = 0.6893$. Next, we choose the largest $100(1 - \hat{\pi})\%$ of the data and use their average as the signal amplitude $\hat{\mu} = 3.194$. The observations in later stages will be generated based on the estimated parameters.

We set both the FDR and MDR at level 0.1, apply SMART and record the total sample size. We then apply DS with the recorded sample size by SMART. The results are summarized below. Since DS always eliminates half of the locations at each step, for this particular instance DS requires at least $1.5p$ observations, and does not offer proper error rate control. Next, we run DS up to 10 stages and record the FPR and MDR levels, and then apply
SMART at the same levels and compare the required sample sizes. The results are summarized below. We can see that SMART has smaller error rates while taking significantly fewer samples.

| Methods | FDP   | MDP   | Total Observation |
|---------|-------|-------|-------------------|
| DS      | 0.9971195 | 0    | 77641             |
| SMART   | 0.1487284  | 0.02162162 | 67850             |
| SMART   | 0.083333   | 0.1081081   | 56926             |

6. Proofs. This section provides the proofs of all theorems in the main text. The derivation of the approximation formulae (2.8) is provided in the Supplementary Material.

6.1. Proof of Theorem 1.

Proof. Part (a) Denote \( d^x(t_l, t_u) = \{(N_i, \delta_i) : 1 \leq i \leq p\} \). Using the definition of \( T_{OR}^{i, N_i} \), we have

\[
\mathbb{E} \left\{ \sum_{i=1}^{p} (1 - \theta_i) \delta_i \right\} = \mathbb{E}_{\theta_i | X} \left\{ \sum_{i=1}^{p} (1 - \theta_i) \delta_i \right\} = \mathbb{E} \left( \sum_{i=1}^{p} T_{OR}^{i, N_i} \delta_i \right). 
\]

Then the FPR is \( Q(t_l, t_u) = \mathbb{E} \left( \sum_{i=1}^{p} T_{OR}^{i, N_i} \delta_i \right) / \mathbb{E} (\sum_{i=1}^{p} \delta_i) \). It follows that

\[
\mathbb{E} \left[ \sum_{i=1}^{p} \left( T_{OR}^{i, N_i} - Q(t_l, t_u) \right) \mathbb{I} \left( T_{OR}^{i, N_i} \leq t_l \right) \right] = \mathbb{E} \left[ \sum_{i:t_{OR}^{i, N_i} \leq t_l} \left( T_{OR}^{i, N_i} - Q(t_l, t_u) \right) \right] = 0.
\]

The above equation implies that \( Q(t_l, t_u) \leq t_l \); otherwise every term on the LHS must be negative, resulting in a contradiction.

According to Assumption 1, \( \mathbb{P}_{\theta_i} (N_i < \infty) = 1 \) for all \( i \); see [4] for a proof. Since \( \mathbb{P}_{\theta_i} (N_i < \infty) = 1 \) for all \( i \), and \( p \) is finite, we claim that \( \mathbb{P}(\max N_i < \infty) = 1 \), i.e. the oracle procedure has a finite stopping time.

Next we prove that for a fixed \( t_u \), \( Q_{OR}(t_l, t_u) \) is non-decreasing in \( t_l \). Let \( Q_{OR}(t_{l,j}, t_u) = \alpha_j \) for \( j = 1, 2 \). We only need to show that if \( t_{l,1} < t_{l,2} \), then \( \alpha_1 \leq \alpha_2 \). Denote \( N_{i,1} \) and \( N_{i,2} \) the stopping times for location \( i \) corresponding to thresholds \( (t_{l,1}, t_u) \) and \( (t_{l,2}, t_u) \), respectively. If \( t_{l,1} < t_{l,2} \), then it is easy to see that for any particular realization of the experiment, we must have \( N_{i,1} \geq N_{i,2} \).
We shall show that if \( t_{l,1} < t_{l,2} \) and \( \alpha_1 > \alpha_2 \), then we will have a contradiction. To see this, note that

\[
\left( T_{OR}^{i,N_i,2} - \alpha_2 \right) \mathbb{I} \left( T_{OR}^{i,N_i,2} \leq t_{l,2} \right) = \mathbb{E} \left\{ \sum_{i=1}^{p} \left( T_{OR}^{i,N_i,2} - \alpha_2 \right) \mathbb{I} \left( T_{OR}^{i,N_i,2} \leq t_{l,2} \right) \right\} - \mathbb{E} \left\{ \sum_{i=1}^{p} T_{OR}^{i,N_i,1} \mathbb{I} \left( T_{OR}^{i,N_i,1} \leq t_{l,1} \right) \right\} = 0.
\]

The second equality holds because if \( T_{OR}^{i,N_i,2} < t_{l,1} \), then we must have \( N_{i,1} = N_{i,2} \). Taking expectations on both sides, we have

\[
\mathbb{E} \left\{ \sum_{i=1}^{p} T_{OR}^{i,N_i,1} \mathbb{I} \left( T_{OR}^{i,N_i,1} \leq t_{l,1} \right) \right\} = 0.
\]

However, since \( \alpha_1 > \alpha_2 \) and \( \alpha_1 \leq t_{l,1} \) as shown previously, we must have

\[
\mathbb{E} \left\{ \sum_{i=1}^{p} \left( \alpha_1 - \alpha_2 \right) \mathbb{I} \left( T_{OR}^{i,N_i,1} \leq t_{l,1} \right) \right\} = 0.
\]

This leads to a contradiction. Therefore, we conclude that \( Q_{OR}(t_l, t_u) \) is non-decreasing in \( t_l \) for a fixed \( t_u \).

Next, we prove that \( \tilde{Q}(t_l, t_u) \) is non-increasing in \( t_u \) for a fixed \( t_l \). By the definition of MDR and similar arguments for the FPR part, we have

\[
\mathbb{E} \left[ \sum_{i=1}^{p} \left( 1 - T_{OR}^{i,N_i} \right) \mathbb{I} \left( T_{OR}^{i,N_i} \geq t_u \right) - \tilde{Q}(t_l, t_u) \right] = 0.
\]

Since our model has a finite stopping time, naturally we have \( \{ i : T_{OR}^{i,N_i} \geq t_u \} \cup \{ j : T_{OR}^{j,N_j} \leq t_l \} = \{ 1, 2, 3, \ldots, p \} \). It follows that

\[
\mathbb{E} \left[ \sum_{i:T_{OR}^{i,N_i} \geq t_u} \left( 1 - T_{OR}^{i,N_i} \right) \left( 1 - \tilde{Q}(t_l, t_u) \right) \right] = \mathbb{E} \left\{ \sum_{j:T_{OR}^{j,N_j} \leq t_l} \left( 1 - T_{OR}^{j,N_j} \right) \tilde{Q}(t_l, t_u) \right\}.
\]
We have
\[
\frac{1 - \tilde{Q}(t_l, t_u)}{\tilde{Q}(t_l, t_u)} = \frac{\mathbb{E} \left\{ \sum_{j : T_{\text{OR}}^{i,N_i} \leq t_u} (1 - T_{\text{OR}}^{i,N_i}) \right\}}{\mathbb{E} \left\{ \sum_{i : T_{\text{OR}}^{i,N_i} \geq t_u} (1 - T_{\text{OR}}^{i,N_i}) \right\}}.
\]

Consider two thresholds \( t_{u,1} > t_{u,2} \). Denote \( N_{i,1} \) and \( N_{i,2} \) the corresponding stopping times at location \( i \). The operation of the thresholding procedure implies that \( N_{i,1} \geq N_{i,2} \), \( \{ i : T_{\text{OR}}^{i,N_i} \geq t_{u,1} \} \subset \{ i : T_{\text{OR}}^{i,N_i} \geq t_{u,2} \} \), and \( \{ j : T_{\text{OR}}^{j,N_j,2} \leq t_l \} \subset \{ j : T_{\text{OR}}^{j,N_j,1} \leq t_l \} \). Therefore, we have
\[
\mathbb{E} \left\{ \sum_{i : T_{\text{OR}}^{i,N_i} \geq t_{u,2}} (1 - T_{\text{OR}}^{i,N_i}) \right\} = \mathbb{E} \left\{ \sum_{i : T_{\text{OR}}^{i,N_i} \geq t_{u,1}} (1 - T_{\text{OR}}^{i,N_i}) \right\} + \mathbb{E} \left\{ \sum_{i : T_{\text{OR}}^{i,N_i} \geq t_{u,2}} (1 - T_{\text{OR}}^{i,N_i}) \right\} \geq \mathbb{E} \left\{ \sum_{i : T_{\text{OR}}^{i,N_i} \geq t_{u,1}} (1 - T_{\text{OR}}^{i,N_i}) \right\}.
\]

We have shown that \( \{ j : T_{\text{OR}}^{j,N_j,2} \leq t_l \} \subset \{ j : T_{\text{OR}}^{j,N_j,1} \leq t_l \} \). Moreover, on the set \( \{ j : T_{\text{OR}}^{j,N_j,2} \leq t_l \} \), we have \( N_{i,1} = N_{i,2} \). It follows that
\[
\mathbb{E} \left\{ \sum_{j : T_{\text{OR}}^{j,N_j,1} \leq t_l} (1 - T_{\text{OR}}^{j,N_j,1}) \right\} \geq \mathbb{E} \left\{ \sum_{j : T_{\text{OR}}^{j,N_j,2} \leq t_l} (1 - T_{\text{OR}}^{j,N_j,2}) \right\}.
\]

Combining the above results, we have
\[
\frac{1 - \tilde{Q}(t_l, t_u)}{\tilde{Q}(t_l, t_u)} \geq \frac{1 - \tilde{Q}(t_l, t_{u,2})}{\tilde{Q}(t_l, t_{u,2})}.
\]

Hence if \( t_{u,1} > t_{u,2} \), then it follows from (6.1) that
\[
\frac{1 - \tilde{Q}(t_l, t_{u,1})}{\tilde{Q}(t_l, t_{u,1})} \geq \frac{1 - \tilde{Q}(t_l, t_{u,2})}{\tilde{Q}(t_l, t_{u,2})}.
\]

Therefore \( \tilde{Q}(t_l, t_{u,1}) \leq \tilde{Q}(t_l, t_{u,2}) \). We conclude that \( \tilde{Q}(t_l, t_u) \) is non-increasing in \( t_u \) for a fixed \( t_l \).
Part (b). The proof is divided into two parts. The first part describes a process that identifies a unique pair of \((t_{OR}^l, t_{OR}^u)\). The second part shows that \(d^\pi(t_{OR}^l, t_{OR}^u)\) has the largest power among all eligible procedures.

(1). Oracle thresholds. Let \(Q(1,1) = \tilde{\alpha}\) be the theoretical upper bound corresponding to the FPR when all hypotheses are rejected. A prespecified FPR level \(\alpha > 0\) is called eligible if \(\alpha < \tilde{\alpha}\). Let \(\mathcal{R}_\alpha = \{t_u : Q(t_u, t_u) > \alpha\}\). We can see that \(\mathcal{R}_\alpha\) is nonempty if \(\alpha\) is eligible, since \(Q(0,0) = 0\) and \(Q(1,1) = \tilde{\alpha}\). Consider \(t_u \in \mathcal{R}_\alpha\). Note that \(Q(0, t_u) = 0\) for all \(t_u\), the following threshold is well defined:

\[
(6.2) \quad t_{OR}^l(t_u) = \sup\{t_l : Q(t_l, t_u) \leq \alpha\}.
\]

We claim that \(Q\{t_{OR}^l(t_u), t_u\} = \alpha\).

We prove by contradiction. First, according to the continuity of \(Q(t_l, t_u)\), for every \(t_u \in \mathcal{R}_\alpha\), we can find \(t_u^*\) such that \(Q(t_u^*, t_u) = \alpha\) [since \(Q(0, t_u) = 0\) and \(Q(t_u, t_u) > \alpha\)]. If not the equality does not hold, i.e. we have \(Q\{t_{OR}^l(t_u), t_u\} < \alpha\), then the monotonicity of \(Q(t_l, t_u)\) implies that \(t_u^* > t_{OR}^l(t_u)\), which contradicts the definition of \(t_{OR}^l(t_u)\). The above construction shows that, for every \(t_u \in \mathcal{R}_\alpha\), we can always identify a unique \(t_{OR}^l(t_u)\) such that \(Q\{t_{OR}^l(t_u), t_u\} = \alpha\).

We say \((\alpha, \gamma)\) constitute an eligible pair of prespecified error rates if \(\alpha\) is eligible, and for this \(\alpha, \gamma\) satisfies \(0 < \gamma < \sup\left\{\tilde{Q} \left( t_{OR}^l(t_u), t_u \right) : t_u \in \mathcal{R}_\alpha \right\}\). In the above definition, the eligibility of \((\alpha, \gamma)\) only depends on the model, but not any given \(t_u\). Now consider an eligible pair \((\alpha, \gamma)\). The continuity of \(Q(t_u) \equiv \tilde{Q}\{t_{OR}^l(t_u), t_u\}\) implies that we can find \(t_u^*\) such that \(\tilde{Q}\{t_{OR}^l(t_u^*), t_u^*\} = \gamma\). Let \(t_{OR}^u = \inf\{t_u \in \mathcal{R}_\alpha : \tilde{Q}(t_{OR}^l(t_u), t_u) = \gamma\}\). The pair of oracle thresholds are thus given by \((t_{OR}^l, t_{OR}^u) \equiv (t_{OR}^l(t_u^*), t_{OR}^u(t_u^*))\).

(2). Proof of optimality. Denote \(d_* = (N_*, \delta_*)\) a sequential procedure that satisfies FPR\((d_*) = \alpha_* \leq \alpha\), MDR\((d_*) = \gamma_* \leq \gamma\), where \(N_* = (N_*^1, \ldots, N_*^p)\) and \(\delta_* = (\delta_*^1, \ldots, \delta_*^p)\) are the corresponding stopping times and decision rules. Deonte EAST\((d_*)\) the expected average stopping times. By definition, we have

\[
\mathbb{E}\left\{\sum_{i=1}^p (T_{OR}^{i, N_*^i} - \alpha_* \delta_*^i)\right\} = 0, \quad \mathbb{E}\left\{\sum_{i=1}^p (1 - T_{OR}^{i, N_*^i})(1 - \delta_*^i - \gamma_*)\right\} = 0.
\]

Now we can sort \(T_{OR}^{i, N_*^i}\) as \(T_{OR}^{(1), N_*^{(1)}} \leq T_{OR}^{(2), N_*^{(2)}} \leq \cdots \leq T_{OR}^{(p), N_*^{(p)}}\) with their corresponding decisions \(\delta_*^{(1)}, \delta_*^{(2)}, \cdots, \delta_*^{(p)}\). If \(\delta_*\) does not take the form of

\[
(6.3) \quad \text{there exists a } k, \text{ such that } \delta_*^{(i)} = \begin{cases} 1 & i \leq k \\ 0 & k < i \leq p \end{cases},
\]
then we can always modify $\delta^*$ into such a form with the same EAST and smaller FPR and MDR. Specifically, suppose that there exists $l_1 < l_2$ such that $\delta^{(l_1)} = 0$ and $\delta^{(l_2)} = 1$, then we swap these two decisions. Such operation can be iterated until the decision rule takes the form as (6.3). Denote the new decision rule by $d'_s = (N_s, \delta'_s)$. Since $T^{(l_1)}_{OR}, N^{(l_1)}_i \leq T^{(l_2)}_{OR}, N^{(l_2)}_i$ in each swapping, we can reduce the FPR and MDR:

$$\alpha'_s = \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} \leq \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} = \alpha_s,$$

$$\gamma'_s = \frac{\sum_{i=1}^p \left( 1 - T^{i,N_i}_{OR} (1 - \delta^i_s) \right)}{p \pi} \leq \frac{\sum_{i=1}^p \left( (1 - T^{i,N_i}_{OR}) (1 - \delta^i_s) \right)}{p \pi} = \gamma_s.$$

Expressing $d'_s$ in the form of (2.5), we can find $t'_l$ and $t'_u$ such that $\delta_{t,N_t} = 1$ for a fixed $t$, we must have $\nabla_t = 1$, then we swap these two decisions. Such operation can be iterated until the decision rule takes the form of (6.3). Denote the new decision rule by $d'_s = (N_s, \delta'_s)$. Since $T^{(l_1)}_{OR}, N^{(l_1)}_i \leq T^{(l_2)}_{OR}, N^{(l_2)}_i$ in each swapping, we can reduce the FPR and MDR:

$$\alpha'_s = \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} \leq \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} = \alpha_s,$$

$$\gamma'_s = \frac{\sum_{i=1}^p \left( 1 - T^{i,N_i}_{OR} (1 - \delta^i_s) \right)}{p \pi} \leq \frac{\sum_{i=1}^p \left( (1 - T^{i,N_i}_{OR}) (1 - \delta^i_s) \right)}{p \pi} = \gamma_s.$$

Expressing $d'_s$ in the form of (2.5), we can find $t'_l$ and $t'_u$ such that $\delta_{t,N_t} = 1$ for a fixed $t$, we must have $\nabla_t = 1$, then we swap these two decisions. Such operation can be iterated until the decision rule takes the form of (6.3). Denote the new decision rule by $d'_s = (N_s, \delta'_s)$. Since $T^{(l_1)}_{OR}, N^{(l_1)}_i \leq T^{(l_2)}_{OR}, N^{(l_2)}_i$ in each swapping, we can reduce the FPR and MDR:

$$\alpha'_s = \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} \leq \frac{\sum_{i=1}^p \left( T^{i,N_i}_{OR} \delta^i_s \right)}{\mathbb{E} \left( \sum_{i=1}^p \delta^i_s \right)} = \alpha_s,$$

$$\gamma'_s = \frac{\sum_{i=1}^p \left( 1 - T^{i,N_i}_{OR} (1 - \delta^i_s) \right)}{p \pi} \leq \frac{\sum_{i=1}^p \left( (1 - T^{i,N_i}_{OR}) (1 - \delta^i_s) \right)}{p \pi} = \gamma_s.$$

Next, assume that $t'_l > t^{u}_{OR}(t^{u}_{OR})$. Then by the definition of $t^{u}_{OR}(t^{u}_{OR})$ and monotonicity of $Q(t_l,t_u)$, we have $Q(t'_l,t'_u) = \gamma$. The fact that $Q(t'_l,t'_u) > \gamma$, contradicting the fact that $Q(t'_l,t'_u) = \alpha'_s < \alpha$. Therefore we must have $t'_u \geq t^{u}_{OR}$. Hence we must have $t'_l \geq t^{u}_{OR}(t^{u}_{OR})$. Therefore $\text{EAST}(d'_s) = \text{EAST}(d'_s) \geq \text{EAST}(d_{OR})$ and the desired result follows.

6.2. Proof of Theorem 2.

Proof. The goal is to show that the pair $t^{l}_{OR} = \alpha$ and $t^{u}_{OR} = \frac{1-\pi}{\pi \gamma + 1-\pi}$ control the FPR and MDR. The FPR part is straightforward since

$$\text{FPR}(\hat{d}_{OR}) = \frac{\mathbb{E} \left\{ \sum_{i=1}^p T^{i,N_i}_{OR} I(T^{i,N_i}_{OR} \leq \alpha) \right\}}{\mathbb{E} \left\{ \sum_{i=1}^p I(T^{i,N_i}_{OR} \leq \alpha) \right\}} \leq \alpha.$$
To show the MDR part, we first carry out an analysis of the false negative rate (FNR), which is defined as 
\[
\text{FNR}(\tilde{d}_{OR}) = \frac{\mathbb{E}\left\{ \sum_{i=1}^{p} (1 - \theta_i)(1 - \delta_i) \right\}}{\mathbb{E}\left\{ \sum_{i=1}^{p} (1 - \delta_i) \right\}}.
\]
According to the operation of \(\tilde{d}_{OR}\), the FNR can be further calculated as
\[
\text{FNR}(\tilde{d}_{OR}) = \frac{\mathbb{E}\left\{ \sum_{i=1}^{p} (1 - \tilde{T}_{i,OR} \delta_i) \right\}}{\mathbb{E}\left\{ \sum_{i=1}^{p} (1 - \delta_i) \right\}} \leq 1 - t_u = \frac{\pi \gamma}{\pi \gamma + 1 - \pi}.
\]

Denote \(\tilde{d}_{OR} = (\tilde{N}_{OR}, \tilde{\delta}_{OR})\). We have shown that \(\text{FPR}(\tilde{d}_{OR}) \leq \alpha\). Suppose the actual FPR level is \(\tilde{\alpha} \leq \alpha\). Then
\[
\mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i (1 - \tilde{\delta}_i) \right\} = \tilde{\alpha} \mathbb{E}\left\{ \sum_{i=1}^{p} \tilde{\delta}_i \right\}.
\]
\[
\text{Meanwhile, our analysis of the FNR shows that}
\]
\[
\text{(6.4)} \quad (1 - \tilde{\alpha}) \mathbb{E}\left\{ \sum_{i=1}^{p} \tilde{\delta}_i \right\} = \mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \tilde{\delta}_i \right\}.
\]
Combining (6.4) and (6.5), we obtain
\[
(\pi \gamma + 1 - \pi) \left\{ p \pi - (1 - \tilde{\alpha}) \mathbb{E}\left\{ \sum_{i=1}^{p} \tilde{\delta}_i \right\} \right\} \leq \pi \gamma \left\{ p - \mathbb{E}\left\{ \sum_{i=1}^{p} \tilde{\delta}_i \right\} \right\}.
\]
It follows that \(\mathbb{E}\left\{ \sum_{i=1}^{p} \tilde{\delta}_i \right\} \geq \frac{p \pi (1 - \pi)(1 - \gamma)}{\pi \gamma (1 - \pi)(1 - \tilde{\alpha})} \geq \frac{p \pi (1 - \gamma)}{1 - \tilde{\alpha}}\). Using (6.4) and \(p \pi = \mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \right\}\), we have \(\mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \tilde{\delta}_i \right\} \geq \mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \right\} - \gamma \mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \right\}\). Therefore, \(\mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i (1 - \tilde{\delta}_i) \right\} \leq \gamma \mathbb{E}\left\{ \sum_{i=1}^{p} \theta_i \right\}\) and the desired result follows.

6.3. Proof of Theorem 3.

**Proof.** Define stage-wise false positive rate \(s\text{FPR}_j\) and stage-wise false non-discovery rate \(s\text{FNR}_j\) as
\[
s\text{FPR}_j := \frac{\mathbb{E}\left\{ \sum_{i \in S_j} (1 - \theta_i) \delta_i \right\}}{\mathbb{E}\left\{ \sum_{i \in S_j} \delta_i \right\}}, \quad s\text{FNR}_j := \frac{\mathbb{E}\left\{ \sum_{i \in S_j} \theta_i (1 - \delta_i) \right\}}{\mathbb{E}\left\{ \sum_{i \in S_j} (1 - \delta_i) \right\}},
\]
where \(s\text{FPR}_j\) is the ratio of the expected number of false rejections at stage \(j\) over the expected number of all rejections at stage \(j\), and \(s\text{FNR}_j\) is the ratio...
MULTISTAGE ADAPTIVE TESTING

of the expected number of false acceptance over the expected number of all acceptance. By our definition of \( sFPR_j \), we have

\[
sFPR_j = \frac{\mathbb{E} \left( \sum_{i=1}^{k_j} T_{i,j}^{(i)} \right)}{\mathbb{E} \left( k_j \right)} \leq \frac{\mathbb{E} \left( k_j \alpha \right)}{\mathbb{E} \left( k_j \right)} = \alpha.
\]

Similarly, by the definition of \( sFNR_j \), we have

\[
sFNR_j = \frac{\mathbb{E} \left\{ \sum_{i=0}^{k_j-1} \left( 1 - T_{i,j} \right) \right\}}{\mathbb{E} \left( k_j \right)} \leq \frac{\mathbb{E} \left( k_j \frac{\pi \gamma}{\pi \gamma + 1 - \pi} \right)}{\mathbb{E} \left( k_j \right)} = \frac{\pi \gamma}{\pi \gamma + 1 - \pi}.
\]

Therefore SMART controls the sFDR and sFNR at level \( \alpha \) and \( \frac{\pi \gamma}{\pi \gamma + 1 - \pi} \), respectively, across all stages.

Next we show that if \( sFPR_j \) and \( sFNR_j \) are controlled universally at pre-specified levels across all stages, then the global FPR and FNR can be controlled at the same levels. Consider the global FPR first:

\[
FPR = \frac{\mathbb{E} \left\{ \sum_{j=1}^{N} \sum_{i \in S_j} (1 - \theta_i) \delta_{SM}^i \right\}}{\mathbb{E} \left\{ \sum_{j=1}^{N} \sum_{i \in S_j} \delta_{SM}^i \right\}} \leq \frac{\mathbb{E} \left( \sum_{j=1}^{N} \alpha \sum_{i \in S_j} \delta_{SM}^i \right)}{\mathbb{E} \left( \sum_{j=1}^{N} \sum_{i \in S_j} \delta_{SM}^i \right)} = \alpha.
\]

Next, we consider the global FNR:

\[
FNR = \frac{\mathbb{E} \left\{ \sum_{j=1}^{N} \sum_{i \in S_j} \theta_i \left( 1 - \delta_{SM}^i \right) \right\}}{\mathbb{E} \left\{ \sum_{j=1}^{N} \sum_{i \in S_j} \left( 1 - \delta_{SM}^i \right) \right\}} \leq \frac{\mathbb{E} \left\{ \sum_{j=1}^{N} \frac{\pi \gamma}{\pi \gamma + 1 - \pi} \sum_{i \in S_j} \left( 1 - \delta_{SM}^i \right) \right\}}{\mathbb{E} \left\{ \sum_{j=1}^{N} \sum_{i \in S_j} \left( 1 - \delta_{SM}^i \right) \right\}} = \frac{\pi \gamma}{\pi \gamma + 1 - \pi}.
\]

Finally, according to the arguments in Theorem 2, the MDR satisfies \( MDR \leq \gamma \) if \( FNR \leq \frac{\pi \gamma}{\pi \gamma + 1 - \pi} \), completing the proof.

6.4. Proof of Theorem 4.

Proof. For a symmetric decision procedure \( d \), denote its Type I and Type II errors on unit \( i \) by \( \alpha' = \mathbb{P}_{H_{i,0}}(\text{Reject } H_{i,0}) \) and \( \gamma' = \mathbb{P}_{H_{i,1}}(\text{Accept } H_{i,0}) \). It can be shown that the corresponding global error rates are given by

\[
(6.6) \quad \text{FPR}(d) = \frac{(1 - \pi)\alpha'}{(1 - \pi)\alpha' + \pi(1 - \gamma')} \quad \text{and} \quad \text{MDR}(d) = \gamma',
\]

\[
\]
respectively. Our result largely follows from the lower bound derived in [24] on family-wise error rate (FWER); we only highlight the main steps on how to go from the FWER paradigm to the FPR/MDR paradigm, which essentially involves exploiting the relationship (6.6). From Thm. 2.39 in [35], we have

\[
\tau_1 \geq \frac{\alpha' \log (\frac{\alpha'}{1-\gamma'}) + (1-\alpha') \log (\frac{1-\alpha'}{\gamma'})}{D(F_0||F_1)}
\]

\[
\tau_2 \geq \frac{(1-\gamma') \log (\frac{1-\gamma'}{\alpha'}) + \gamma' \log (\frac{\gamma'}{1-\alpha'})}{D(F_1||F_0)},
\]

where \(\tau_1\) and \(\tau_2\) are the expected stopping times for null and non-null locations, respectively. Furthermore, from [24], we have

\[
\eta \geq \eta' \geq (\frac{1}{1-\pi}) \log (2^{\frac{1}{\alpha'}}),
\]

Using the KL divergence \(D_{KL}(F_0, F_1) = \max \{D(F_0||F_1), D(F_1||F_0)\}\), the average stopping time of all locations satisfies

\[
\tau = \frac{(p-p\pi)\tau_1 + p\pi\tau_2}{p} \geq \frac{(1-\pi)(1-\alpha') \log (\gamma')^{-1} + \pi(1-\gamma') \log (\alpha')^{-1} - \log 2}{D_{KL}(F_0, F_1)}.
\]

(6.7) We consider two situations. If \(\alpha' \leq \gamma'\), then

\[
\tau \geq (1-\pi)(1-\gamma') \log (\gamma')^{-1} + \pi(1-\gamma') \log (\gamma')^{-1} - \log 2.
\]

Note that for \(0 \leq x \leq 1\), \(\{x \log x : x \in (0, 1)\}\) reaches its minimum when \(x = e^{-1}\). It follows that \(2^{-1} < e^{-1/e} \leq \gamma' \leq 1\). Therefore \((1-\gamma') \log (\gamma')^{-1} \geq \log(2\gamma')^{-1}\). Together with (6.7), we have \(\tau \geq \frac{\log (4\gamma')^{-1}}{D_{KL}(F_0, F_1)}\). According to our constraint on \(\tau\), we conclude that \(\log (4\eta)^{-1} \geq \tau D_{KL}(F_0, F_1) \geq \log (4\gamma')^{-1}\). Therefore, \(\gamma' \geq \eta\) and \(R^*(d) \geq \eta\).

If \(\gamma' \leq \alpha'\), then we can similarly show that

\[
\tau \geq \frac{(1-\alpha') \log \alpha'^{-1} - \log 2}{D_{KL}(F_0, F_1)} \geq \frac{\log \frac{1}{2\alpha'} - \log 2}{D_{KL}(F_0, F_1)} \geq \frac{\log \frac{1}{4\alpha'}}{D_{KL}(F_0, F_1)}.
\]

It follows that \(\log (4\eta)^{-1} \geq \tau D_{KL}(F_0, F_1) \geq \log (4\alpha')^{-1}\), which implies \(\alpha' \geq \eta\).

Under the assumption that \(\pi < \frac{1}{3}\) and \(\eta \leq \frac{1}{2}\), we have \(\eta \leq \frac{1}{2} \leq \frac{1-2\pi}{1-\pi}\). Consider
the function $k(x) = \frac{(1-\pi)x}{(1-\pi)x + \pi}$. It is easy to see that $k(x)$ is monotonically increasing in $x$. Hence

$$R^*(d) \geq \frac{(1-\pi)\alpha'}{(1-\pi)\alpha' + \pi(1-\gamma')} \geq \frac{(1-\pi)\alpha'}{(1-\pi)\eta + \pi} \geq \frac{(1-\pi)(1-2\pi)}{1-\pi + \pi} = \eta,$$

completing the proof.

6.5. Proof of Theorem 5.

Proof. We have already shown that when $t_l = \alpha$ and $t_u = \frac{1-\pi}{\pi \gamma + 1-\pi}$, the SMART procedure controls the FDR and MDR at level $\alpha$ and $\gamma$, respectively. Let $\alpha = \gamma = \frac{1}{f(p) + \epsilon}$. With the choice of $t_l$ and $t_u$ mentioned above, we have

$$\lim_{p \to \infty} R^*(d) = \lim_{p \to \infty} \frac{2}{f(p)^{1+\epsilon}} = 0,$$

which proves the first part of the theorem.

Next we establish the upper bound. Consider $p$ simultaneous SPRTs with the same threshold $t_l$ and $t_u$. The operation of our SMART procedure uses these thresholds for the moving averages; hence the SPRT approach with the same $t_l$ and $t_u$ will always take more samples. It is sufficient to show that the result holds for simultaneous SPRTs.

We convert thresholds $t_l$ and $t_u$ to the thresholds for SPRTs: $A = \frac{(1-\pi)(1-t_u)}{\pi t_u}$, $B = \frac{(1-\pi)(1-t_l)}{\pi t_l}$. Under our specifications, we further have $A = \frac{\gamma(1-\pi)}{\pi \gamma + 1-\pi}$, $B = \frac{(1-\pi)(1-\alpha)}{\pi \alpha}$. From [24], we have $\alpha' \leq B^{-1}$, $\gamma' \leq A$. According to Assumption (3.1), we have

$$\mathbb{E}(\log L_{i,N_i}^i | \log L_{i,N_i}^i < \log A) \geq \log A - C_1,$$

$$\mathbb{E}(\log L_{i,N_i}^i | \log L_{i,N_i}^i > \log B) \leq \log B + C_2$$

for some positives constants $C_1$ and $C_2$. Consider $\mathbb{E}_{\theta_i=0}(\log L_{i,N_i}^i)$. Then

$$-\mathbb{E}_{\theta_i=0}(\log L_{i,N_i}^i) = -(1-\alpha)\mathbb{E}_{\theta_i=0}(\log L_{i,N_i}^i | \log L_{i,N_i}^i < \log A) - \alpha \mathbb{E}_{\theta_i=0}(\log L_{i,N_i}^i | \log L_{i,N_i}^i > \log B) \leq (1-\alpha)(\log A^{-1} + C_1) - \alpha \log B \leq (1-\alpha)(\log A^{-1} + C_1) \leq \log A^{-1} + C_1.$$
Likewise, we can show that $E_{\theta_i=1}(\log L_{i,N_i}) \leq \log B + C_2$. Let $C_3 = (1 - \pi)C_1 + \pi C_2$. According to Wald’s identity (P490, [4]), we have

\[
E_{\theta_i=1}(N_i) = \frac{E_{\theta_i=1}(\log L_{i,N_i})}{D(F_1|F_0)} = \frac{E_{\theta_i=1}(\log L_{i,N_i})}{D(F_1|F_0)},
\]

\[
E_{\theta_i=0}(N_i) = \frac{E_{\theta_i=0}(\log L_{i,N_i})}{D(F_0|F_1)} = \frac{-E_{\theta_i=0}(\log L_{i,N_i})}{D(F_0|F_1)}.
\]

It follows that

\[
\lim_{p \to \infty} \frac{\tau}{\log f(p)} = \lim_{p \to \infty} \frac{(1 - \pi)E_{\theta_i=0}(N_i) + \pi E_{\theta_i=1}(N_i)}{\log f(p)} 
\leq \lim_{p \to \infty} \frac{(1 - \pi)(\log A^{-1} + C_1)}{\log f(p)D(F_0|F_1)} + \lim_{p \to \infty} \frac{\pi(\log B + C_2)}{\log f(p)D(F_1|F_0)} 
\]

\[
= \lim_{p \to \infty} \frac{(1 - \pi)\log \frac{\pi \gamma + 1 - \pi}{\gamma (1 - \pi)} + \pi \log \frac{(1 - \pi)(1 - \alpha)}{\pi \alpha} + C_3}{\log f(p) \min \{(D(F_0|F_1), D(F_1|F_0))\}} 
\]

\[
= \lim_{p \to \infty} \frac{(1 - \pi)\log \gamma^{-1} + \pi \log \frac{(1 - \pi)}{\pi \alpha} + C_3}{\log f(p) \min \{(D(F_0|F_1), D(F_1|F_0))\}} 
\]

\[
= \lim_{p \to \infty} \frac{\log(\alpha^{-1})}{\min \{(D(F_0|F_1), D(F_1|F_0))\}} 
\]

\[
= \frac{1 + \epsilon}{\min \{(D(F_0|F_1), D(F_1|F_0))\}}.
\]

From (6.8) to (6.9), we have used the fact $\alpha$ and $\gamma$ are error rates converging to zero; hence $\frac{\pi \gamma + 1 - \pi}{\gamma (1 - \pi)} = \frac{1}{\gamma} \{1 + o(1)\}$ and $\frac{(1 - \pi)(1 - \alpha)}{\pi \alpha} = (1 - \pi)\{1 + o(1)\}$. Equation (6.10) uses the fact that $\alpha = \gamma$. The desired result follows. \qed

References.

[1] BASHAN, E., NEWSTADT, G. and HERO, A. O. (2011). Two-stage multiscale search for sparse targets. IEEE Transactions on Signal Processing 59 2331–2341.

[2] BENJAMINI, Y. and HELPER, R. (2007). False discovery rates for spatial signals. J. Amer. Statist. Assoc. 102 1272–1281. MR2412549

[3] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B 57 289–300. MR1325392 (96d:62143)

[4] BERGER, J. O. (1985). Statistical decision theory and Bayesian analysis. Springer.

[5] BLANCHARD, G. and GEMAN, D. (2005). Hierarchical Testing Designs for Pattern Recognition. Ann. Statist. 33 1155–1202.

[6] BLEICHER, K. H., BÖHM, H.-J., Müller, K. and ALANINE, A. I. (2003). Hit and lead generation: beyond high-throughput screening. Nature reviews Drug discovery 2 369–378.
[7] Bloma, P. E., Fleischerb, S. J. and Smilowitzb, Z. (2002). Spatial and Temporal Dynamics of Colorado Potato Beetle in Fields with Perimeter and Spatially Targeted Insecticides. *Environmental Entomology* **31** 149–159.

[8] Cai, T. T., Jin, J. and Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653 (2009c:62058)

[9] Copas, J. B. (1974). On symmetric compound decision rules for dichotomies. *Ann. Statist.* **2** 199–204.

[10] Cormack, R. (1988). Statistical challenges in the environmental sciences: a personal view. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **201–210.

[11] Djorgovski, S. G., Mahabal, A., Donalek, C., Graham, M. J., Drake, A. J., Moghaddam, B. and Turmon, M. (2012). Flashes in a star stream: Automated classification of astronomical transient events. In *E-Science (e-Science), 2012 IEEE 8th International Conference on* 1–8. IEEE.

[12] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 (2005e:62066)

[13] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289

[14] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B* **64** 499–517. MR1924303 (2003h:62027)

[15] Ghosh, M. (1960). Bounds for the expected sample size in a sequential probability ratio test. *Journal of the Royal Statistical Society. Series B (Methodological)* **360–367.

[16] Goeman, J. J. and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24** 537–544.

[17] Goeman, J. J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* **38** 3782–3810.

[18] Haupt, J., Castro, R. M. and Nowak, R. (2011). Distilled Sensing: Adaptive Sampling for Sparse Detection and Estimation. *IEEE T. Inform. Theory* **57** 6222–6235.

[19] Haupt, J., Baraniuk, R. G., Castro, R. M. and Nowak, R. D. (2012). Sequentially designed compressed sensing. In *SSP* 401–404.

[20] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802.

[21] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* **65–70.

[22] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

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

[24] Malloy, M. L. and Nowak, R. D. (2014a). Sequential testing for sparse recovery. *IEEE Transactions on Information Theory* **60** 7862–7873.

[25] Malloy, M. L. and Nowak, R. D. (2014b). Near-optimal adaptive compressed sensing. *IEEE Transactions on Information Theory* **60** 4001–4012.

[26] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393.

[27] Meinshausen, N., Bickel, P., Rice, J. et al. (2009). Efficient blind search: Optimal power of detection under computational cost constraints. *The Annals of Applied Statistics* **3** 38–60.

[28] Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American
[29] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950 131–148. University of California Press, Berkeley and Los Angeles. MR0044803 (13,480d)

[30] Roesch, F. A. (1993). Adaptive cluster sampling for forest inventories. Forest Science 39 655–669.

[31] Rossell, D. and Müller, P. (2013). Sequential stopping for high-throughput experiments. Biostatistics 14 75–86.

[32] Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J. D., Real, F. X., Van Den Berg, D., Matullo, G., Baris, D. et al. (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nature genetics 42 978–984.

[33] Sarkar, S. K. (2004). FDR-controlling stepwise procedures and their false negatives rates. J. Stat. Plan. Infer. 125 119–137.

[34] Satagopan, J. M., Venkatraman, E. and Begg, C. B. (2004). Two-Stage Designs for Gene–Disease Association Studies with Sample Size Constraints. Biometrics 60 589–597.

[35] Siegmund, D. (1985). Sequential analysis: tests and confidence intervals. Springer Science & Business Media.

[36] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. J. Amer. Statist. Assoc. 102 901–912. MR2411657

[37] Sun, W. and Wei, Z. (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. Biometrics 102 267–280.

[38] Taylor, J., Tibshirani, R. and Efron, B. (2005). The “missed rate” for the analysis of gene expression data. Biostatistics 6 111–117.

[39] Thompson, S. K. and Seber, G. A. (1996). Adaptive sampling. John Wiley & Sons, Inc., New York, NY.

[40] Tony Cai, T. and Sun, W. (2017). Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79 197–223.

[41] Wald, A. (1945). Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics 16 117–186.

[42] Wald, A. (1947). Sequential analysis. Wiley, New York.

[43] Wei, D. and Hero, A. O. (2013). Multistage adaptive estimation of sparse signals. IEEE Journal of Selected Topics in Signal Processing 7 783–796.

[44] Wei, D. and Hero, A. O. (2015). Performance guarantees for adaptive estimation of sparse signals. IEEE Transactions on Information Theory 61 2043–2059.

[45] Yekutieli, D. (2008). Hierarchical False Discovery Rate–Controlling Methodology. Journal of the American Statistical Association 103 309–316.

[46] Zhang, J.-H., Chung, T. D. and Oldenburg, K. R. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. Journal of biomolecular screening 4 67–73.
Supplementary Material for “Multistage Adaptive Testing of Sparse Signals”

APPENDIX A: DERIVATION OF THRESHOLDS

We need the following assumption in our derivation. The assumption has been commonly adopted in the literature on SPRT (e.g. Berger, 1985; Siegmund, 1985).

ASSUMPTION 2. Consider $Z_{i,1}$ defined in Assumption 1. For all $i$, we have $P_{\theta_i}(Z_{i,1} = 0) < 1$, $P_{\theta_i}(|Z_{i,1}| < \infty) = 1$, $P_{\theta_i}(Z_{i,1} < 0) > 0$, and $P_{\theta_i}(Z_{i,1} > 0) > 0$. Moreover, $M_{\theta_i}(t) = E_{\theta_i}[e^{tZ_{i,1}}]$ exists for all $t$.

We start our derivation by noting that $T_{OR}^{i,j}$ is a monotone function of the likelihood ratio statistic $L_{i,j}$:

\[
T_{OR}^{i,j} = P\left( \theta_i = 0 \mid X_j \right) = 1/\left( 1 + \frac{\pi}{1 - \pi} L_{i,j} \right).
\]

Hence $d^\theta(t_i, t_u)$ can be expressed as a thresholding rule based on $L_{i,j}$:

- stops sampling for unit $i$ at time $N_i = \min\{ j \geq 1 : L_{i,j} \leq A \text{ or } L_{i,j} \geq B \}$,
- deciding $\delta_{i,N_i} = 0$ if $L_{i,j} \leq A$ and $\delta_{i,N_i} = 1$ if $L_{i,j} \geq B$.

We first solve $(A, B)$ for a given pair $(\alpha, \gamma)$, then transform $(A, B)$ to $(t_i, t_u)$. The technique used in our derivation is similar to the classical ideas when deriving the upper and lower thresholds for SPRT. Since all testing units operate independently and have the same thresholds, it is sufficient to focus on the operation of SPRT on a generic testing unit. Hence, for simplicity, we drop index $i$ and denote $L_{N}, \theta, L_{i,k}$ as $L_N, \theta$ and $Z_{i,k}$, respectively.

Under the random mixture model, the FPR and MDR of the SPRT with thresholds $(A, B)$ can be calculated as

\[
\text{FPR} = \frac{(1 - \pi)P(L_N > B \mid \theta = 0)}{P(L_N > B)}, \quad \text{MDR} = P(L_N < A \mid \theta = 1).
\]

Let $S_N = \sum_{k=1}^{N} Z_{i,k} = \log L_N$. Denote $a = \log A$ and $b = \log B$. Under Assumption 2, $P_{\theta_i}(N < \infty) = 1$ and all moment of $N$ exist. There exists a unique nonzero number $t_\theta$ for which $M_{\theta_i}(t_\theta) = 1$ ([4]). This fundamental identity then implies

\[
1 = E_{\theta_i} \{ \exp(t_\theta S_N)M_{\theta_i}(t_\theta)^{-N} \} = E_{\theta_i} \{ \exp(t_\theta S_N) \} \\
\approx \exp(t_\theta a)P_{\theta_i}(S_N \leq a) + \exp(t_\theta b)P_{\theta_i}(S_N \geq b).
\]
In the above approximation, we ignore the overshoots and pretend that $S_N$ hits the boundaries $a$ and $b$ exactly. In this idealized situation, $S_N$ has a two-point distribution $P_{\theta}^*: P_{\theta}^*(S_N = a) = P_{\theta}(S_N \leq a)$ and $P_{\theta}^*(S_N = b) = P_{\theta}(S_N \geq b)$.

Moreover, Assumption 2 implies that

$$1 = P_{\theta}(N < \infty) = P_{\theta}(L^N \leq A) + P_{\theta}(L^N \geq B)$$
$$= P_{\theta}(S_N \leq a) + P_{\theta}(S_N \geq b).$$

Thus we can solve from the above that

$$P_{\theta}(L^N \geq B) = P_{\theta}(S_N \geq b) \approx \left\{ -\exp(t_{\theta}a) \right\} / \left\{ \exp(t_{\theta}b) - \exp(t_{\theta}a) \right\}.$$ 

According to Assumption 2, $P_{\theta}(|Z_{-k}| < \infty) = 1$ for $\theta = 0, 1$. Then $t_{\theta=0} = 1, t_{\theta=1} = -1$ (P493, [4]). It follows that

$$\text{FPR} \approx \frac{(1 - \pi) \frac{1-A}{B-A}}{P(L^N \geq B)} = \frac{(1 - \pi) \frac{1-A}{B-A}}{(1 - \pi) \frac{1-A}{B-A} + \frac{1-1/A}{1/B-1/A}} = 1 - \pi / (1 - \pi + \pi B),$$
$$\text{MDR} \approx 1 - \frac{1 - \exp(-a)}{\exp(-b) - \exp(-a)} = 1 - \frac{1 - 1/A}{1/B-1/A} = \frac{A(B-1)}{B-A}.$$ 

Setting $\text{FPR} = \alpha, \text{MDR} = \gamma$ and solving for $A$ and $B$, we have

$$A \approx \frac{(\alpha^{-1} - 1)(1 - \pi)\gamma}{(\alpha^{-1} - 1)(1 - \pi) - \pi + \pi \gamma}, \quad B \approx \frac{(\alpha^{-1} - 1)(1 - \pi)}{\pi}.$$ 

The relationship (A.11) implies $A = (1 - \pi)(1 - t_u)/(\pi t_u), \quad B = (1 - \pi)(1 - t_l)/(\pi t_l)$. Transforming from $L^{i,j}$ to $T_{OR}^{i,j}$, the corresponding thresholds can be obtained as:

$$t_{OR}^l = \alpha \quad \text{and} \quad t_{OR}^u = \frac{\pi \alpha \gamma + 1 - \pi - \alpha}{\pi \gamma + 1 - \pi - \alpha}.$$ 

To ensure an effective MDR control, we choose a more stringent upper threshold:

$$t_{OR}^u = \frac{1 - \pi}{\pi \gamma + 1 - \pi} \geq \frac{\pi \alpha \gamma + 1 - \pi - \alpha}{\pi \gamma + 1 - \pi - \alpha}, \quad \forall \alpha \geq 0.$$ 

**APPENDIX B: APPLICATION TO ASTRONOMICAL SURVEY**

In astronomical surveys, a common goal is to separate sparse targets of interest (stars, supernovas, or galaxies) from background noise. We consider a dataset from Phoenix Deep Survey (PDS), a multi-wavelength survey of a region over 2 degrees diameter in the southern constellation Phoenix. Fig. 3(a) shows a telescope image from the PDS. It has $616 \times 536 = 330,176$ pixels,
among which 1131 pixels exhibit signal amplitude of at least 2.98. In practice we monitor the same region for a fixed period of time. After taking high resolution images, it is of interest to narrow down the focus quickly using a sequential testing procedure so that we can use limited computational resources to explore certain regions more closely. The image is converted into gray-scale with signal amplitudes standardized. Fig. 3(b) depicts a contaminated image with Gaussian white noise.

We apply SMART by setting both the FPR and MDR at 5%, then record the total number of measurements, and finally apply DS with the recorded sample size. We can see that SMART control the error rates precisely. The resulting images for SMART and DS are demonstrated in Fig. 3(c) and (d), respectively. We can see SMART produces much sharper images than DS.

| Methods | FDP       | MDP       | Total Observation |
|---------|-----------|-----------|-------------------|
| SMART   | 0.06321335| 0.05658709| 368796           |
| DS      | 0.9864338 | 0.00265252| 495866           |

Next we implement DS up to 12 stages, record the FDR and MDR levels, and then apply SMART at the recorded error rates. The required sample sizes of the two methods are summarized below. We can see that SMART control the error rates with fewer observations. The resulting images for SMART and DS are shown in Fig. 3(e) and (f), we can see SMART produces slightly sharper images than DS.

| Methods | FDP       | MDP       | Total Observation |
|---------|-----------|-----------|-------------------|
| DS      | 0.07237937| 0.01414677| 670331           |
| SMART   | 0.03192407| 0.00795756| 479214           |
Fig 3: SMART and DS comparison. Fig. (a) and (b) show the original radio telescope image and tainted image with white noise, respectively. Fig. (c) and (d) compare SMART and DS when the total number of observations are about the same. Fig. (f) shows the resulting image when implementing DS for 12 stages, Fig. (e) shows the image produced by SMART when using the recorded error rates from the 12-stage DS.