Empirical Bayes Multistage Testing for Large-Scale Experiments

Hui Xu, Weinan Wang

September 14, 2022

Abstract

Modern application of A/B tests is challenging due to its large scale in various dimensions, which demands flexibility to deal with multiple testing sequentially. The state-of-the-art practice first reduces the observed data stream to always-valid p-values, and then chooses a cut-off as in conventional multiple testing schemes. Here we propose an alternative method called AMSET (adaptive multistage empirical Bayes test) by incorporating historical data in decision-making to achieve efficiency gains while retaining marginal false discovery rate (mFDR) control that is immune to peeking. We also show that a fully data-driven estimation in AMSET performs robustly to various simulation and real data settings at a large mobile app social network company.

Keywords: Always valid inference; Multiple testing; Sequential A/B testing
1 Introduction

Large-scale online experiments (George et al., 2005; Gerber and Green, 2012) are routinely conducted at various technology companies (Google, Netflix, Microsoft, LinkedIn, etc.) (Xie and Aurisset, 2016; Kohavi et al., 2009, 2013; Tang et al., 2010; Bakshy et al., 2014; Xu et al., 2015) to help gauge quantifiable impacts of potential feature and product launches. At a large mobile app social network company, hundreds of experiments are conducted on a daily basis, aiming to improve user engagement or app-performance. To accurately measure these changes, hundreds and thousands of metrics are further pumped through our in house A/B platform where routine statistical tests are conducted to judge if the treatment effects are indeed statistically significant.

Given such high velocity in experiments, we are able to gather a rich repository of historical data related to metric sensitivities and treatment effects, which could potentially be team and product specific. However, traditional p-value based decision making that’s calculated via two-sample t-tests fail to leverage such data, since the test statistic relies only on a single experiment’s single metric data. This inefficiency not only translates to loss in statistical power (Sun and Cai, 2007) in treatment effect detection, but also limits the type of decision making available to us that is otherwise possible from an empirical Bayesian perspective.

Furthermore, although best practices inform experiment practitioners to hold off until the end to examine experiment efficacy, oftentimes people ”hunt” for significance (p-value < 0.05) throughout the whole experiment duration. The incentive for continuously monitoring experimental results is strong since detecting true effects and giving up null effects as early as possible reduces the opportunity cost from running longer experiments, especially when estimating optimal run-time in advance is difficult. However, if no corrections are made, continuous monitoring would augment the type I error incurred, as demonstrated in (Johari et al., 2015). Due to invalid p-values under continuous monitoring, false discovery rate (FDR) control is also compromised.

Combining these two pivotal needs, efficient leverage of historical data and continuous monitoring without the pain of type I error inflation, we propose an empirical Bayes procedure named AMSET (adaptive multistage empirical Bayes test), together with a fully data-driven estimation procedure that is shown to perform robustly to various simulation and real data settings. Comparing with the state-of-the-art method of always valid p-value utilized by Optimizely (Johari et al., 2021), our method AMSET can control mFDR level at arbitrary stopping time with better precision, and also shows significant power gains.

1.1 Main contributions

Our main contribution is two-folds. First, in the oracle situation where population level parameters are known, we derived a Bayes optimal decision rule based on a weighted classification problem and showed that it minimizes the number of missed discoveries subject to marginal false discovery rate (mFDR) control under mild regularity conditions at any time stage. Our second contribution is that we developed an adaptive multistage empirical Bayes (AMSET) procedure by approximating the Bayes optimal decision rule with compound thresholding together with a fully data-driven estimation scheme. We proved that the oracle AMSET gives valid marginal false discovery rate (mFDR) control at any stopping time, thus allowing continuous monitoring of A/B test platforms. In addition, we demonstrated significant power gain of AMSET at mFDR level α as compared to the state-of-the-art always valid p-value approach in various simulation and real data scenarios for both the oracle and data-driven settings.
1.2 Organization of the paper

The remainder of the paper is organized as follows. In section 2, we present related literature. In section 3, we describe our setup and notations. In section 4, we establish optimality of the chosen test statistic in fixed horizon setting and introduce our proposed AMSET procedure with known population level parameters, which gives valid mFDR control at any stopping time. In section 5, we describe a fully data-driven procedure that estimates population level parameters using empirical Bayes methods. In section 6, we do a brief recap of the state-of-the-art method of always-valid p-value and clarify connections to our test statistics. In section 7, we compare our method in both the oracle and data-driven scenario with the method used by Optimzely in simulations of fixed-horizon and stagewise testing. In section 8, we demonstrate our method in a real data example using data from large scale A/B testing platform.

2 Background and Related Work

Multiple hypothesis testing can be viewed as a compound decision problem in the sequential context, where each component is testing of a single hypothesis based on sequentially obtained observations.

2.1 Sequential hypothesis testing

Sequential A/B test framework was initially formulated by Wald in 1945 and was widely used in pharmaceutical clinical trials. It is known that the sequential probability ratio test (SPRT) is optimal for the single simple vs simple hypothesis testing problem, such that it minimizes the expected number of observations among all sequential procedures with given levels of type I and type II errors. For a comprehensive understanding of related theory, see Berger (2013) and Siegmund (2013) for details. Adoption of existing sequential tests in online A/B test platforms is limited since validity for arbitrary stopping time is not guaranteed.

There has been recent interest in constructing valid sequential tests that allow for continuous monitoring. In Johari et al. (2017), Johari et al. (2021), the authors proposed a general framework of always valid p-values and suggested construction using a particular family of sequential tests called mixture sequential probability ratio tests (mSPRT) Robbins (1970), which are test of power one Robbins and Siegmund (1974) and is asymptotically optimal if the cost per experiments diminishes to zero Lai (2001). In Wasserman et al. (2020), split likelihood-ratio test (split LRT) was introduced and extended to yield always valid p-values and confidence sequences.

It is worth mentioning that there are other goals besides any time validity that are of interest in sequential testing. For example, constructing sequences of intervals that contain an unknown parameter with uniformly high probability over an infinite time horizon with Law of the Iterated Logarithm Darling and Robbins (1967), Balsubramani and Ramdas (2015)), and using sequential A/B testing to find the best alternative with multi-armed bandits Bubeck et al. (2012), Lattimore and Szepesvári (2020), Scott (2015).

2.2 Multiple hypothesis testing

Starting from multiple testing in the single-stage design, the standard procedure to control false discovery rate (FDR) is Benjamini-Hochberg Benjamini and Hochberg (1995), abbreviated as BH.
Sun and Cai proposed an adaptive z-value based procedure \cite{sun2007} that is optimal in the sense that it controls mFDR at level $\alpha$ with the smallest mFNR, where mFNR is the proportion of the expected number of nonnulls among the expected number of nonrejections.

Due to the feasibility of conducting large-scale online experiments on a routine basis, there has also been recent interest in achieving multiple hypothesis testing controls in the sequential contexts. \cite{wang2017} built on \cite{sun2007} and constructed a simultaneous multistage adaptive ranking and thresholding procedure that mimics the optimality of SPRT, i.e. minimizes the expected total number of samples while satisfying pre-specified constraints on both false positive rate (FPR) and missed discovery rate (MDR). Similarly as SPRT, the procedure is not compatible with continuous monitoring.

The only state-of-the-art method that provides always valid FDR control in sequential multiple hypothesis testing is to combine BH procedure with always valid p-values \cite{johari2017, johari2021}. While there are other efforts to obtain always valid multiple testing control in the sequential context, works in this area consider a different data stream from the one we focus on in this paper. For example, a class of methods called “$\alpha$-spending” and “$\alpha$-investing” was used to control the family wise error rate (FWER) or the false discovery rate (FDR), where a different hypothesis is evaluated at each time stage \cite{demets1994, foster2008, javanmard2018}. There are also works that provide alternative multi-armed bandit (MAB) frameworks for controlling false discoveries when multiple experiments are run over time. In \cite{yang2017}, sequence of fixed A/B tests are replaced with best-arm MAB instances. In \cite{jamieson2018}, traffic is adaptively assigned to treatment with the aim of minimizing sample size to reach target true positive proportion, while ensuring anytime false discovery control. Besides, there are also works that focus on combining continuous monitoring with other metrics beyond FDR \cite{shi2022, malek2017, shi2021, gronau2019}.

In sequential testing, it may not be necessary to measure all coordinates for all time stage, especially when running long experiments are costly. Adaptive sampling can be used in the sense that the data sampling procedure varies across coordinates. A plethora of powerful multistage testing and estimation have been developed under this more flexible data sampling regime, such as hierarchical testing procedures for pattern recognition \cite{blanchard2005, meinshausen2009, sun2015}, distilled sensing and sequential thresholding methods for sparse detection \cite{haupt2011, malloy2011, malloy2014}, etc. $\alpha$–investing can be considered as an extreme case of adaptive sampling where each hypothesis is considered once. In this paper, we will focus on a particular type of adaptive sampling where the set of hypothesis to sample from decreases or filters down at each time stage.

\section{Setup and Notations}

Consider the well established problem \cite{sun2007, jin2007, wang2017, tony2019} of recovering the support of a high dimensional vector $\theta = (\theta_1, \ldots, \theta_m) \in \{0, 1\}^m$ from i.i.d. sequential measurements of variables $X_1, \ldots, X_m$. At each time stage $t = 1, 2, \ldots$, we observe an independent measurement of each variable,

$$X_{it} \sim (1 - \theta_i)F_0 + \theta_i F_i,$$  \hspace{1cm} (1)
for $i = 1, \ldots, m$. That is to say, $X_{it}$ follows null distribution $F_0$ if $\theta_i = 0$ and the non-null distribution $F_i$ if $\theta_i = 1$. In the context of large-scale A/B testing, $F_0$ is the distribution of two-sample t-statistics with mean zero approximated by standard normal and $F_i$’s are approximately normal distributions with different means corresponding to effect sizes.

Our goal is to develop a cost effective and (asymptotically) valid decision rule to identify the support of $\theta$ (i.e. non-zero treatment effects) reliably in the sequential testing setting. To be more precise, let $\delta_t = (\delta_{t1}, \ldots, \delta_{tm})$ denote the decision rule based on $F_t = \sigma(\{x_{is}\}_{i \in [m], s \leq t})$, where $\delta_{ti} \in \{0, 1\}$ indicates whether coordinate $i$ is classified as a null or non-null case. For any fixed time stage $t$, define False Discovery Rate (FDR) as $FDR(\delta_t) = \mathbb{E}[FDP(\delta_t)]$, where

$$FDP(\delta_t) = \frac{\sum_{i=1}^{m} (1 - \theta_i) \delta_{ti}}{(\sum_{i=1}^{m} \delta_{ti}) \lor 1}. \quad (2)$$

An asymptotically equivalent measure is the marginal false discovery rate (mFDR) (Genovese and Wasserman (2002)) defined as

$$mFDR(\delta_t) = \frac{\mathbb{E}[\sum_{i=1}^{m} (1 - \theta_i) \delta_{ti}]}{\mathbb{E}[(\sum_{i=1}^{m} \delta_{ti}) \lor 1]} . \quad (3)$$

While mFDR or FDR guarantees validity, cost effectiveness can be measured by missed discovery rate (MDR),

$$MDR(\delta_t) = \mathbb{E} \left[ \frac{\sum_{i=1}^{m} \theta_i (1 - \delta_{ti})}{(\sum_{i=1}^{m} \theta_i) \lor 1} \right]. \quad (4)$$

The goal of large-scale A/B testing can therefore be formulated as minimizing MDR subject to FDR (or mFDR) control. For any stopping time $\tau$, we can similarly define $FDP(\delta^\tau)$, $mFDR(\delta^\tau)$, and $mFDR(\delta^\tau)$, where the expectation is with respect to all randomness including stopping time $\tau$. 

Figure 1: Illustration of AMSET procedure. The left part of the diagram corresponds to multistage data and mixture deconvolution obtained either by oracle or data-driven estimations. In real settings, alternative density $f_1$ could be Gaussian mixtures.
4 Sequential A/B Test

We first consider an ideal situation where an oracle knows all population level parameters including the proportion of non-nulls \( p \), as well as the densities of null \( f_0 \) and non-null \( f_1 \).

4.1 Optimal sequential test for fixed horizon

We start by restricting to a non-adaptive setting, in the sense that data for each coordinate is collected at all time stages. Suppose that \( \theta_i \) are iid Bernoulli random variables with \( \mathbb{P}(\theta_i = 1) = p \).

Define the following oracle test statistic for stage \( T \) and coordinate \( i \),

\[
T^T_{\text{OR}} = \frac{(1 - p) \prod_{t=1}^{T} f_0(x_{it})}{(1 - p) \prod_{t=1}^{T} f_0(x_{it}) + p \prod_{t=1}^{T} f_1(x_{it})}.
\]

(5)

Since we assume that the \( x_{it} \) are independent draws from the same distribution for all time stages, the oracle test statistic is the local false discovery rate (lfdr), \( \mathbb{P}(\theta_i = 0 | x_i, 1:T) \) \((Efron et al. (2001))\), given the data stream up to time stage \( T \). Truncation with lfdr was found to be the optimal test in many situations of both sequential testing \((Berger (2013))\) and multiple testing \((Sun and Cai (2007))\).

For any time stage \( T \) and mFDR level of \( \alpha \in (0, 1) \), we define a weighted classification problem with the following loss function

\[
L_{\lambda}(\theta, \delta^T) = \frac{1}{m} \left[ \sum_{i=1}^{m} \lambda (1 - \theta_i) \delta^T_i + \theta_i (1 - \delta^T_i) - \alpha \lambda \delta^T_i \right],
\]

(6)

where \( \lambda > 0 \) is the relative weight for false positive. The following theorem shows that \( T^T_{\text{OR}} \) minimizes the classification risk \( \mathbb{E}[L_{\lambda}(\theta, \delta^T)] \) at any time stage \( T \).

**Theorem 1.** Suppose that \( \theta_i \), \( i = 1, \ldots, m \), are iid Bernoulli(\( p \)) random variables, and \( x_{it} \) are independent draws from the mixture model in equation (1) for all \( t = 1, \ldots, T \). Given known values of \( p, f_0, f_1 \), the classification risk \( \mathbb{E}[L_{\lambda}(\theta, \delta^T)] \) is minimized by \( \delta^B(\lambda, T) = (\delta^B_1(\lambda, T), \ldots, \delta^B_m(\lambda, T)) \) for any \( T \geq 1 \), where

\[
\delta^B_i(\lambda, T) = 1\{T^T_{\text{OR}} \leq \frac{1 + \alpha \lambda}{1 + \lambda}\}.
\]

(7)

The above theorem shows that the Bayes optimal sequential test for the weighted classification problem is a truncation rule based on lfdr statistic. We then make connections with the optimal solution in the large-scale A/B testing problem. For the sake of notation, let \( G^T_0 \) and \( G^T_1 \) be the cumulative distribution functions (cdf) of \( T^T_{\text{OR}} \) under the null and the alternative, respectively, then the marginal distribution for \( T^T_{\text{OR}} \) can be written as \( (1 - p)G^T_0 + pG^T_1 \).

**Theorem 2.** Under the same data generating process as in Theorem 1. Let \( 1 - p > \alpha \). Assume that \( G^T_0 \) and \( G^T_1 \) are continuous functions and

\[
\frac{(1 - p)G^T_0(\alpha)}{(1 - p)G^T_0(\alpha) + pG^T_1(\alpha)} < \alpha,
\]

(8)

then there exists a \( \lambda^* > 0 \) such that the solution to the weighted classification problem in equation (7) is also optimal in the multiple testing problem at time stage \( T \) in the sense that, among all
decision rules at time $T$ that control mFDR at level $\alpha$, it has the smallest expected proportion of missed discoveries. That is, $\delta^B(\lambda^*, T)$ is the solution to the constraint optimization problem after observing data until stage $T$.

\[
\min_{\{\delta_i\}_{i=1}^m} \frac{1}{m} \sum_{i=1}^m \mathbb{E}[\theta_i(1 - \delta_i)] \quad \text{s.t.} \quad \text{mFDR}(\delta) = \frac{\mathbb{E}[\sum_{i=1}^m (1 - \theta_i)\delta_i]}{\mathbb{E}[\sum_{i=1}^m \delta_i] \lor 1} \leq \alpha.
\] (9)

The above theorem establishes that truncation of lfdr is also optimal in the multiple testing problem at any fixed time stage $T$. For large-scale A/B testing, the signal is usually sparse, where the proportion of non-nulls $p$ is usually in the range of $(0.01, 0.1)$, thus satisfying the weak assumption $1 - p > \alpha$.

The assumption of (8) is equivalent to the statement that mFDR of simple truncation rule at level $\alpha$ is less than $\alpha$.

**Proposition 1.** Under the same data generating process as in Theorem 2. For any coordinate $i$ and fixed time stage $T$,

\[
\frac{(1 - p)G_0^T(\alpha)}{(1 - p)G_0^T(\alpha) + pG_1^T(\alpha)} \leq \alpha.
\] (10)

From the above proposition, we know that assumption (8) only requires the inequality to be strict. That is, the simple thresholding rule is more conservative than the nominal level $\alpha$. In other words, there is additional gain by increasing the truncation threshold without overshooting in terms of mFDR.

We have motivated the choice of oracle test statistic by establishing optimality of lfdr-based test in fixed time (horizon) setting. Given observed data until fixed time stage $T$, the decision rule $\delta^B(\lambda^*, T)$ minimizes missed discovery rate among all tests that control mFDR at level $\alpha$. Since power is $1 - MDR$, it is equivalent to maximizing power subject to mFDR control.

### 4.2 Adaptive sequential test

Instead of sampling for every coordinate at each time stage, we may want to consider adaptive sampling, in order to achieve some balance between power and sampling cost. While we have shown in the previous section that the Bayes optimal test at every time stage is a thresholding function based on observations at all past time stages, it might suffer from sample inefficiency, which is undesirable for monetization A/B tests. For this purpose we consider the following adaptive procedure

Stop sampling and reject for coordinate $i$ at time stage

$N_i = \min \left\{ t \geq 1 : T_{OR}^i \leq \frac{1 + \alpha \lambda_t}{1 + \lambda_t} \right\}$,

where $\lambda_t$ could vary across time points. In other words, we maintain an active set $A_t = \{ i : N_i \geq t \}$ for sampling and testing at each time stage $t$. 

6
The adaptive sampling procedure allows us to save cost from data collection and further computation of test statistic. Adaptive sampling will not compromise global mFDR control, and the thresholding rule based on our oracle test statistic is stage-wise optimal (conditioning on stagewise active set) as shown in Lemma 1.

### 4.3 Oracle AMSET

In this section, we will introduce the method of AMSET (adaptive multistage empirical Bayes test) that combines the oracle test statistics motivated in 4.1 and the adaptive procedure described in 4.2 to arrive at a sequential procedure for multiple testing that has valid mFDR control at any stopping time, and thus is immune to peeking.

Let $A_t$ be the active set at stage $t$. Starting with the full active set $A_1 = \{1, \ldots, m\}$ for all $i \in \{1, \ldots, m\}$. We truncate the lfdr test statistics using compound thresholding (Sun and Cai (2007)) and update the active set at every time stage $t$ for a pre-specified level $\alpha > 0$ with details listed in Algorithm 1 and a diagram illustration in Figure 1. The non-adaptive version that we call multistage empirical Bayes test (MSET) in Algorithm 2 makes decisions based on full observations at each coordinate and all time stages and therefore has higher power. In general we recommend AMSET over MSET for monetization A/B tests in order to save experimentation cost as a tradeoff for power.

**Algorithm 1: Oracle AMSET**

Set $\alpha$ and start with full active set $A_1 = \{1, \ldots, m\}$;
Specify any stopping criterion;
for time stages $t = 1, 2, \ldots$ do

- **(Data Sampling)** Observe data $\{x_{it}\}_{i \in A_t}$;
- **(Ranking)** For all $i \in A_t$, compute $T_{OR}^{(i,t)}$ in Eq (5) based on $\{x_{is}\}_{s \leq t}$ and sort them in ascending order
  
  $T_{OR}^{(1),t}, T_{OR}^{(2),t}, \ldots, T_{OR}^{(k_t),t},$

  where $k_t = \text{Card}(A_t)$;
- **(Compound Thresholding)** Define threshold
  
  $r_t = \max \left\{ r : \frac{1}{r} \sum_{i=1}^{r} T_{OR}^{(i),t} \leq \alpha \right\}$

  Let $S_t = \{ i \in A_t : T_{OR}^{(i,t)} \leq T_{OR}^{(r),t} \}$;
- **(Updating)** Let $A_{t+1} = A_t \setminus S_t$;
- **(Decision)** Let $\delta^t_i = 1$ for all $i \notin A_{t+1}$ and 0 otherwise;
  Break if stopping criterion is satisfied or $A_{t+1} = \emptyset$.

The procedure AMSET offers flexibility in terms of cost saving since there is no need to collect data once a specific coordinate is rejected. That is to say, a rejected hypothesis will always remain rejected. AMSET also allows us the convenience to continuously monitor the sequential testing results with mFDR guarantee at any stopping time as stated in the following Theorem.
**Theorem 3.** For any stopping time $\tau$, let $\delta^\tau = (\delta^\tau_1, \ldots, \delta^\tau_m)$ generated by Algorithm 1. Then

$$\text{mFDR}(\delta^\tau) \leq \alpha.$$  

We also have fixed-horizon mFDR and FDR control at level $\alpha$, i.e.

$$\text{FDR}(\delta^t|S_t) = \mathbb{E} \left[ \frac{\sum_{i \in S_t} (1 - \theta_i) \delta^t_i}{(\sum_{i \in S_t} \delta^t_i) \lor 1} \right] \leq \alpha,$$

and

$$\text{mFDR}(\delta^t|S_t) = \frac{\mathbb{E} \left[ \sum_{i \in S_t} (1 - \theta_i) \delta^t_i \right]}{\mathbb{E} \left[ (\sum_{i \in S_t} \delta^t_i) \lor 1 \right]} \leq \alpha,$$

for any fixed $t$.

**Remark 1.** The compound thresholding step in the procedure above can be replaced by a simple thresholding, that is, we reject all $i \in A_t$ such that $T^t_{\text{OR}} \leq \alpha$. Since all coordinates rejected by simple thresholding will also be rejected by compound thresholding, the former is a more conservative procedure and therefore still satisfies mFDR control at any stopping time.

## 5 Data-driven Implementation

In practice, we do not have the oracle knowledge of true population parameters ($f_1$, $f_0$, and $p$) and need a completely data-driven procedure that can be used for industry applications. Given large scale testing results in historical data, we can harness the power of established empirical Bayes methods in order to estimate the oracle test statistic of $T^t_{\text{OR}}$. For the purpose of A/B testing, the null distribution is approximately $N(0,1)$ due to central limit theorem and the alternative distribution is approximately a mixture of gaussians. The difficulty with of data driven procedure thus lies in the estimation of non-null proportion $p$ and deconvolution of gaussian mixtures using relevant historical data.

1. (Estimate $\hat{p}$) For estimation of the non-null proportion $p$, we use the method proposed in Jin and Cai (2007) based on the empirical characteristic function and Fourier analysis, which is shown to be uniformly consistent over a wide class of parameters.

2. (Estimate $\hat{f}_1$) For estimation of non-null density, we will use Kiefer-Wolfowitz maximum likelihood estimation for Gaussian mixtures with software “REBayes” developed by Koenker and Gu (2017) and computation supported by Koenker and Mizera (2014). The method belongs to the general estimation strategy of g-estimation and has consistency result as stated in Kiefer and Wolfowitz (1956).

Additional details of deconvolution and recovery of $\hat{f}_1$ can be found in the Appendix B.2. Given estimates of non-null proportion $\hat{p}$ and non-null density $\hat{f}_1$, we get the following data-driven test statistic for coordinate $i$ at time stage $T$.

$$T^t_{\text{DD}} = \frac{(1 - \hat{p}) \prod_{t=1}^{T} f_0(x_{it})}{(1 - \hat{p}) \prod_{t=1}^{T} f_0(x_{it}) + \hat{p} \prod_{t=1}^{T} \hat{f}_1(x_{it})},$$

(11)
Algorithm 2: AMSET

Set $\alpha$ and start with full active set $A_1 = \{1, \ldots, m\};$
Specify any stopping criterion;
for time stages $t = 1, 2, \ldots$ do
  (Data Sampling) Observe data $\{x_{it}\}_{i \in A_t};$
  (Ranking) For all $i \in A_t$, compute $T_{DD}^{it}$ in Eq (11) based on $\{x_{is}\}_{s \leq t}$ and sort them in ascending order $T_{DD}^{(1),t}, T_{DD}^{(2),t}, \ldots, T_{DD}^{(k_t),t},$ where $k_t = \text{Card}(A_t);$ (Compound Thresholding) Define threshold
  $$r_t = \max \left\{ r : \frac{1}{r} \sum_{i=1}^{r} T_{DD}^{(i),t} \leq \alpha \right\}$$
Let $S_t = \{i \in A_t : T_{DD}^{it} \leq T_{DD}^{(r_t),t}\}$;
(Updating) Let $A_{t+1} = A_t \setminus S_t;$
(Decision) Let $\delta^*_t = 1$ for all $i \notin A_{t+1}$ and 0 otherwise;
Break if stopping criterion is satisfied or $A_{t+1} = \emptyset.$

6 Comparison with always valid p-value approach

6.1 FDR control with always valid p-value

We compare AMSET (or MSET) with the state-of-the-art method that combines always valid p-values with Benjamini-Hochberg (BH) procedure Benjamini and Hochberg (1995). Recall always valid p-value in Johari et al. (2015) is constructed using mSPRT likelihood ratio test statistic (originally proposed in Robbins and Siegmund (1974)) at stage $T$ for coordinate $i$,

$$\Lambda_T^H = \int_{\Theta} \prod_{t=1}^{T} \frac{f_\theta(x_{it})}{f_{\theta_0}(x_{it})} dH(\theta),$$

where $H(\theta)$ is some prior distribution on parameter $\theta$.

Always valid p-value process for coordinate $i$ is constructed as

$$p_0 = 1, p_T = \min\{p_{T-1}, 1/\Lambda_T^H\},$$

such that for any stopping time $\tau$,

$$\forall s \in [0, 1], P_{\theta_0}(p_\tau \leq s) \leq s.$$

For fair comparison, we use a prior $H$ that has probability 1 on the true mean of alternative
distribution to compare with our oracle procedure, and a prior \( H \sim \mathcal{N}(0, \tau^2) \) where \( \tau^2 = 2 \) to compare with our data driven procedure.

### 6.2 Connection of lfdr to always valid p-values

Both the oracle and data-driven test statistics of lfdr can also be formulated in the framework of always valid p-values using a standard martingale argument for running MLE LRT in universal inference (Wasserman et al. (2020)). For the sake of brevity, we will only illustrate construction of always valid p-value using data-driven statistic \( T^D_{DD} \) for each coordinate \( i \).

Consider the sequential test that rejects the null and stop at time \( T \) if

\[
L_{iT} := \frac{1 - \hat{p} 1 - T^D_{DD}}{T^D_{DD}} \geq 1/\alpha. \tag{12}
\]

Since \( L_{iT} = \frac{\prod_{t=1}^{T} f_1(x_{it})}{\prod_{t=1}^{T} f_0(x_{it})} \) is a nonnegative martingale with respect to the natural filtration, we can use Doob’s inequality to bound the Type-I error as

\[
P_{H_0}(\exists t \geq 1 : L_{it} \geq 1/\alpha) \leq \alpha \mathbb{E}_{H_0}[L_{it}] = \alpha.
\]

This shows that the sequential test defined by Eq (12) is a valid sequential test at level \( \alpha \). Since there is a correspondence between sequential tests and always valid p-values, we know that \( p_{it} = 1/L_{it} \) is an always valid p-value. That is, for any stopping time \( \tau \), \( P_{H_0}(p_{\tau} \leq \alpha) \leq \alpha \). This means that we can also adopt Optimizely’s framework by combining BH procedure with the always valid p-value process constructed with Eq (12) to guarantee finite sample FDR control at any stopping time and thus is immune to peeking.

### 7 Simulation

Throughout our simulations, we compare our proposed method of AMSET with the state-of-the-art method used by Optimizely Johari et al. (2021) that combines standard BH procedure with always valid p-value constructed using mixture sequential probability ratio test (mSPRT). Since AMSET is an adaptive procedure that does not sample from all coordinates at all time stages, there will be possible power loss as compared to the non-adaptive MSET or Optimizely’s always valid p-value approach.

![Figure 2: Setting 1 result for fixed horizon comparison. Fix \( p = 0.05 \), consider \( f_1 \) as the density of \( \mathcal{N}(\mu, 1) \), where \( \mu \) varies from 1 to 3 with grid size of 0.2. Rates are computed as average over 500 simulations.](image-url)
Figure 3: Setting 2 result for fixed horizon comparison. Fix $f_1$ as the density of $N(2, 1)$, vary $p$ from 0.01 to 0.1 with grid size 0.01. Rates are computed as average over 500 simulations.

Figure 4: Setting 3 result for fixed horizon comparison. Vary $p$ from 0.02 to 0.1 with grid size of 0.01 and consider $f_1$ as the density of $N(\mu, 1)$, where $\mu \sim \text{Unif}(2, 4)$ at each location. Rates are computed as average over 500 simulations.

Figure 5: Setting 1 result for stagewise comparison. Fixed $\mu = 2$ for each non-null location and proportion of non-nulls $p = 0.05$. Rates are computed as average over 500 simulations.

For data generation, we consider $m = 5000$ locations with ground truth $\theta_i$ ($1 \leq i \leq m$) drawn i.i.d from Bernoulli distribution with parameter $p$, for some non-null proportion $p \in (0, 1)$. In realistic settings, the signal proportion $p$ is usually small ($\leq 0.1$).

For each time stage, if coordinate $i$ is still in the observation set, we draw an independent sample from $f_0$ if $\theta_i = 0$ and $f_1$ otherwise. In the context of A/B testing, different coordinates corresponds to different randomized experiments and observation $x_{it}$ corresponds to the two-sample t-statistic at time stage $t$ for experiment $i$. Since $f_0$ corresponds to the density of two-sample t-statistic under the null where treatment effect is zero, it can be approximated by standard normal density. Similarly, $f_1$ can be approximated by a Gaussian density with non-zero location $\mu \neq 0$ or a Gaussian mixture with various non-zero locations.

We consider four different tests labeled in the following way. (1) AMSET_OR/MSET_OR: AMSET/MSET with oracle statistic (Eq (5)); (2) AMSET_DD/MSET_DD: AMSET/MSET with data-
driven statistic (Eq. (11)); (3) Optimizely_OR: always valid p-value approach with oracle prior on true alternative; (4) Optimizely_DD: always valid p-value approach with Gaussian prior $\mathcal{N}(0, 2)$. For each setting with a fixed set of hyperparameters $p$ and $f_1$, we will fix mFDR (Eq. (3)) control at nominal level 0.05 and compare FDR (Eq. (2)), as well as MDR (Eq. (4)) across 500 simulations. Since MDR = 1 – power, smaller MDR corresponds to more powerful tests.
We consider simulations in both fixed horizon and stage-wise comparisons regimes. For the fixed horizon setting, we run the simulated sequential tests until a fixed time stage and compare FDR/MDR metrics at the last time stage. This allows us to observe trends for a wide range of hyper-parameter choices. On the other hand, stage-wise comparison illustrates how the trend evolves over time for a chosen set of hyper-parameters.

### 7.1 Fixed horizon comparison

For fixed horizon comparison, we set the number of time stages to be fixed at 5 with the following three scenarios described below.

1. Fix \( p = 0.05 \), consider \( f_1 \) as the density of \( \mathcal{N}(\mu, 1) \), where \( \mu \) varies from 1 to 3 with grid size of 0.2

2. Fix \( f_1 \) as the density of \( \mathcal{N}(2, 1) \), vary \( p \) from 0.01 to 0.1 with grid size 0.01

3. Vary \( p \) from 0.02 to 0.1 with grid size of 0.01 and consider \( f_1 \) as the density of \( \mathcal{N}(\mu, 1) \), where \( \mu \sim \text{Unif}(2, 4) \) at each coordinate

From the results presented in Figure 2-4, it can be observed that both the oracle and data driven versions of the state-of-the-art method used by Optimizely are much more conservative in term of FDR control and has generally lower power than AMSET across all settings and hyperparameter choices. The difference in power diminishes as the signal strength \( \mu \) increases and power goes to 1, but remains approximately the same for different choices of non-null proportions \( p \).

### 7.2 Stagewise comparison

We also use simulation to compare the oracle and data-driven methods of MSET and Optimizely by looking at FDR and MDR at each time stage for a fixed set of hyperparameters as follows.

1. Fixed \( \mu = 2 \) for each non-null location and proportion of non-nulls \( p = 0.05 \)

2. Random \( \mu \sim i.i.d. \text{Unif}(1, 3) \) for each non-null location and proportion of non-nulls \( p = 0.05 \).

From results presented in Figure 5-6, it can be similarly observed that always valid p-value approach used by Optimizely is much more conservative in term of FDR control and has generally lower power than MSET procedure. The difference gets smaller as the number of time stages increases and power for all tests goes to 1.

### 8 Real Data Example

Instead of simulating with arbitrary hyper-parameters, we apply AMSET on actual A/B testing data. That is, we first extract realistic hyper-parameters from historical data of a pool of experiments via Kiefer-Wolfowitz MLE estimation for Gaussian mixtures. And then compare the stagewise performance of various methods under the learnt hyper-parameters.

In particular, we work with the following three metric settings.

1. DIRECT_SEND___total_sent
2. CHAT_CHAT_SEND__chat_video_send

3. TIER1_ACTIVE_DAYS__total_active_days

For each metric, we can aggregate a set of randomized experiments with potential impact on the given metric and compute standardized $z$-scores as estimators for the average treatment effects (ATE). Since each experiment is conducted on a regular basis, we obtain sequences of data for multiple experiments. For example, the CHAT_CHAT_SEND__chat_video_send metric measures the total number of chats sent in the format of videos. Since we have a large repository of sequential A/B tests, in which the metric of interest is measured for both the control and treatment group for three months, the standardized $z$-scores for these tests satisfy our theoretical set up in Section 3, i.e. coming from $F_0$ if ATE = 0, and $F_i$ otherwise for experiment $i$.

| Setting 1 | Setting 2 | Setting 3 |
|-----------|-----------|-----------|
| $\mu$ | Prop. | $\mu$ | Prop. | $\mu$ | Prop. |
| 0 | 0.97 | 0 | 0.98 | 0 | 0.96 |
| 1.61 | 0.03 | 1.32 | 0.01 | 3.36 | 0.01 |
| 1.6 | 0.01 | 17.81 | 0.01 | 19.88 | 0.01 |

By pooling these experiments together, we can learn from historical data and gather useful information on the underlying metric of interest, i.e. the non-null effect size, and the overall pooled proportion of nonzero effects. The estimates of effect sizes with corresponding proportions for each of the three metrics are summarized in the above tables. For data generation, we fix the effect size at each coordinate by drawing from a multinomial distribution specified in the table. Then at each time stage, we sample independently from $\mathcal{N}(\mu, 1)$, where $\mu$ is the specified effect size.

From results presented in Figure 7-9, it can be similarly observed that the always valid p-value based method is much more conservative in term of FDR control and has generally lower power than MSET procedure for all three real data settings. While it has to be acknowledged that the validity of data-driven version of AMSET/MSET is highly dependent on the estimation accuracy of population level parameters, our results show that the data-driven method performs reliably in most cases and yield power gain compared to other methods.

9 Concluding Remarks

The big data era brings tremendous opportunities and also challenges in A/B testing. While enjoying the potential of rich insights generated at unprecedented scale, it is crucial to be aware of the harm of ad-hoc interpretation such as “hunt” for significance. Thus the key is the build a framework that can efficiently uncovers good insights from data, while remaining immune to the common practice of peeking.

This paper proposes an empirical Bayes procedure AMSET for multiple testing under sequential data settings. AMSET is shown to have significant power gains as compared to existing method by leveraging rich amount of historical data available, and at the same time allows continuous monitoring without sacrificing mFDR control.

There are multiple possible future directions based on this work. First there is still possible gap between the compound thresholding rule and the Bayes optimal threshold. Besides, more
exploration in the consistency of data-driven estimation methods can be done to quantify the gap between oracle and data-driven procedures.
References

Bakshy, E., D. Eckles, and M. S. Bernstein (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pp. 283–292.

Balsebramani, A. and A. Ramdas (2015). Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1), 289–300.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Blanchard, G. and D. Geman (2005). Hierarchical testing designs for pattern recognition. *The Annals of Statistics* 33(3), 1155–1202.

Bubeck, S., N. Cesa-Bianchi, et al. (2012). Regret analysis of stochastic and nonstochastic multiarmed bandit problems. *Foundations and Trends® in Machine Learning* 5(1), 1–122.

Darling, D. A. and H. Robbins (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences* 58(1), 66–68.

Demets, D. L. and K. G. Lan (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine* 13(13-14), 1341–1352.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* 32(3), 962–994.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association* 96(456), 1151–1160.

Foster, D. P. and R. A. Stine (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(2), 429–444.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 499–517.

George, E., J. S. Hunter, W. G. Hunter, R. Bins, K. Kirlin IV, and D. Carroll (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley New York.

Gerber, A. S. and D. P. Green (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gronau, Q. F., K. Raj, and E.-J. Wagenmakers (2019). Informed bayesian inference for the a/b test. *arXiv preprint arXiv:1905.02068*.

Haupt, J., R. M. Castro, and R. Nowak (2011). Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory* 57(9), 6222–6235.
Jamieson, K. and L. Jain (2018). A bandit approach to multiple testing with false discovery control. *arXiv preprint arXiv:1809.02235*.

Javanmard, A. and A. Montanari (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics 46*(2), 526–554.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association 102*(478), 495–506.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (2017). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1517–1525.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (2021). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*.

Johari, R., L. Pekelis, and D. J. Walsh (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Koenker, R. and J. Gu (2017). Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software 82*(1), 1–26.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Kohavi, R., T. Crook, and R. Longbotham (2009). Online experimentation at microsoft. *Data Mining Case Studies 11* (2009), 39.

Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1168–1176.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 303–351.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Malek, A., S. Katariya, Y. Chow, and M. Ghavamzadeh (2017). Sequential multiple hypothesis testing with type i error control. In *Artificial Intelligence and Statistics*, pp. 1468–1476. PMLR.

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

Malloy, M. L. and R. D. Nowak (2014). Sequential testing for sparse recovery. *IEEE Transactions on Information Theory 60*(12), 7862–7873.

Meinshausen, N., P. Bickel, and J. Rice (2009). Efficient blind search: Optimal power of detection under computational cost constraints. *The Annals of Applied Statistics 3*(1), 38–60.
Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pp. 131–149. University of California Press.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 41(5), 1397–1409.

Robbins, H. and D. Siegmund (1974). The expected sample size of some tests of power one. *The Annals of Statistics* 2(3), 415–436.

Scott, S. L. (2015). Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry* 31(1), 37–45.

Shi, C., S. Luo, H. Zhu, and R. Song (2021). An online sequential test for qualitative treatment effects. *J. Mach. Learn. Res.* 22, 286–1.

Shi, C., X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song (2022). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 1–13.

Siegmund, D. (2013). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* 102(479), 901–912.

Sun, W. and Z. Wei (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika* 102(2), 267–280.

Tang, D., A. Agarwal, D. O’Brien, and M. Meyer (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 17–26.

Tony Cai, T., W. Sun, and W. Wang (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(2), 187–234.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The annals of mathematical statistics* 16(2), 117–186.

Wang, W. and W. Sun (2017). Sparse recovery with multiple data streams: A sequential adaptive testing approach. *arXiv preprint arXiv:1707.07215*.

Wasserman, L., A. Ramdas, and S. Balakrishnan (2020). Universal inference. *Proceedings of the National Academy of Sciences* 117(29), 16880–16890.

Xie, H. and J. Aurisset (2016). Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 645–654.
Xu, Y., N. Chen, A. Fernandez, O. Sinno, and A. Bhasin (2015). From infrastructure to culture: A/b testing challenges in large scale social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2227–2236.

Yang, F., A. Ramdas, K. Jamieson, and M. J. Wainwright (2017). A framework for multi-a (rmed)/b (andit) testing with online fdr control. arXiv preprint arXiv:1706.05378.
A Proof of results

A.1 Proof of Theorem 1

Proof. By the theorem set up, the posterior distribution of \( \theta_i \) given \( x^{1:T} = \{x_{it}\}_{i \in [m], t \leq T} \) is

\[
P_{\theta_i|x^{1:T}}(\theta_i|x^{1:T}) = \frac{(1-p)(1-\theta_i)\prod_{t=1}^{T} f_{0}(x_{it}) + p\theta_i \prod_{t=1}^{T} f_{1}(x_{it})}{(1-p)\prod_{t=1}^{T} f_{0}(x_{it}) + p\prod_{t=1}^{T} f_{1}(x_{it})}.
\]

Therefore the posterior risk is

\[
E_{\theta|x^{1:T}} L_{\lambda}(\theta, \delta^T) = \frac{1}{m} \sum_{i=1}^{m} \lambda(1-p)\delta^T_i \prod_{t=1}^{T} f_{0}(x_{it}) + p(1-\delta^T_i) \prod_{t=1}^{T} f_{1}(x_{it}) - \alpha\lambda \delta^T_i
\]

\[
= \frac{1}{m} \sum_{i=1}^{m} \left( \frac{p \prod_{t=1}^{T} f_{1}(x_{it})}{(1-p)\prod_{t=1}^{T} f_{0}(x_{it}) + p\prod_{t=1}^{T} f_{1}(x_{it})} \right) + \frac{1}{m} \sum_{i=1}^{m} \left( \frac{(1-\alpha)\lambda(1-p) \prod_{t=1}^{T} f_{0}(x_{it}) - (1+\alpha\lambda)p \prod_{t=1}^{T} f_{1}(x_{it})}{(1-p)\prod_{t=1}^{T} f_{0}(x_{it}) + p\prod_{t=1}^{T} f_{1}(x_{it})} \right) \delta^T_i.
\]

Notice that the posterior risk is minimized by the decision rule

\[
\delta^B_i(\lambda, T) = \left\{ \frac{(1-p) \prod_{t=1}^{T} f_{0}(x_{it})}{p \prod_{t=1}^{T} f_{1}(x_{it})} \leq \frac{1 + \alpha\lambda}{(1-\alpha)\lambda} \right\}
\]

\[
= \left\{ \frac{(1-p) \prod_{t=1}^{T} f_{0}(x_{it})}{(1-p)\prod_{t=1}^{T} f_{0}(x_{it}) + p\prod_{t=1}^{T} f_{1}(x_{it})} \leq \frac{1 + \alpha\lambda}{1 + \lambda} \right\}
\]

This concludes the proof.

A.2 Proof of Theorem 2

Proof. Notice that the constraint of mFDR control can also be written as

\[
E \left[ \sum_{i=1}^{m} (1-\theta_i)\delta_i \right] - \alpha E \left[ \sum_{i=1}^{m} \delta_i \right] \leq 0.
\]

Therefore the Lagrangian dual of the optimization problem in (9) is exactly the classification risk for some \( \lambda > 0 \)

\[
L(\delta, \lambda) = \frac{1}{m} E \left[ \sum_{i=1}^{m} \theta_i (1-\delta_i) + \lambda (1-\theta_i)\delta_i - \lambda\alpha\delta_i \right].
\]  \quad (13)

From Theorem 1 we know that \( \delta^B(\lambda, T) \) is also the optimal solution for the Lagrangian dual problem in (13) for given \( \lambda \). Since \( \lambda > 0 \), we also have \( L(\delta^B(\lambda, T), \lambda) \leq s^* \), where \( s^* \) is the optimal value of
the original constraint optimization problem \([\text{[9]}]\). It remains to show strong duality holds.

We claim that it suffices to show there exists \(\lambda^* > 0\) such that \(\text{mFDR}(\delta^B(\lambda^*, T)) = \alpha\). Assume otherwise, there is some other decision \(\tilde{\delta}\) such that \(\text{mFDR}(\tilde{\delta}) \leq \alpha\) and \(\sum_{i=1}^m E \left[ \theta_i(1 - \tilde{\delta}_i) \right] < \sum_{i=1}^m E \left[ \theta_i(1 - \delta^B_i(\lambda^*, T)) \right]\). Then

\[
\mathcal{L}(\tilde{\delta}, \lambda^*) = \frac{1}{m} \sum_{i=1}^m \theta_i(1 - \tilde{\delta}_i) + \lambda^*(1 - \theta_i)\tilde{\delta}_i - \lambda^* \alpha \tilde{\delta}_i \\
< \frac{1}{m} \sum_{i=1}^m \theta_i(1 - \delta^B_i(\lambda^*, T)) + \lambda^*(1 - \theta_i)\tilde{\delta}_i - \lambda^* \alpha \tilde{\delta}_i \\
\leq \frac{1}{m} \sum_{i=1}^m \theta_i(1 - \delta^B_i(\lambda^*, T)) + \lambda^*(1 - \theta_i)\delta^B_i(\lambda^*, T) - \lambda^* \alpha \delta^B_i(\lambda^*, T) \\
= \mathcal{L}(\delta^B(\lambda^*, T), \lambda^*).
\]

This is a contradiction since \(\delta^B(\lambda^*, T)\) is the minimizer of the Lagrangian with \(\lambda^*\).

We then proceed to show that the constraint equality can be satisfied. First notice that we can write

\[
\text{mFDR}(\delta^B(\lambda, T)) = \frac{(1 - p)G_0^{iT} \left( \frac{1 + \alpha \lambda}{1 + \alpha} \right)}{(1 - p)G_0^{iT} \left( \frac{1 + \alpha \lambda}{1 + \alpha} \right) + pG_1^{iT} \left( \frac{1 + \alpha \lambda}{1 + \alpha} \right)}.
\]

As \(\lambda \to 0\), \(G_0^{iT} \left( \frac{1 + \alpha \lambda}{1 + \alpha} \right), G_1^{iT} \left( \frac{1 + \alpha \lambda}{1 + \alpha} \right) \to 1\). Therefore \(\text{mFDR}(\delta^B(\lambda, T)) \to 1 - p > \alpha\). As \(\lambda \to +\infty\),

\(G_0^{iT} \left( \frac{1 + \alpha \lambda}{1 + \lambda} \right) \to G_0^{iT}(\alpha), G_1^{iT} \left( \frac{1 + \alpha \lambda}{1 + \lambda} \right) \to G_1^{iT}(\alpha)\). Therefore \(\text{mFDR}(\delta^B(\lambda, T)) \to \frac{(1 - p)G_0^{iT}(\alpha)}{(1 - p)G_0^{iT}(\alpha) + pG_1^{iT}(\alpha)} < \alpha\) by assumption. By continuity of \(\text{mFDR}(\delta^B(\lambda, T))\), there exists \(\lambda^* > 0\) such that \(\text{mFDR}(\delta^B(\lambda^*, T)) = \alpha\).

**A.3 Proof of Proposition [1]**

*Proof.* Notice that for the simple truncation decision rule \(\delta^T = (\delta^T_1, \ldots, \delta^T_m)\), where \(\delta^T_i = 1\{T^T_{i, OR} \leq \alpha\},

\[
\text{mFDR}(\delta^T) = \frac{E \left[ \sum_{i=1}^m (1 - \theta_i)\delta^T_i \right]}{E \left[ \sum_{i=1}^m \delta^T_i \lor 1 \right]} = \frac{(1 - p)G_0^{iT}(\alpha)}{(1 - p)G_0^{iT}(\alpha) + pG_1^{iT}(\alpha)}.
\]

For the numerator,

\[
E \left[ \sum_{i=1}^m (1 - \theta_i)\delta^T_i \right] = E \left[ \sum_{i=1}^m (1 - \theta_i)\delta^T_i \mid x_{i,1:T} \right] \\
= E \left[ \sum_{i=1}^m \delta^T_i \mid 1 - \theta_i \mid x_{i,1:T} \right] \\
= E \left[ \sum_{i=1}^m \delta^T_i T_{i, OR} \right] \leq \alpha E \left[ \sum_{i=1}^m \delta^T_i \right],
\]




where the last inequality is due to definition of the truncation rule. Therefore \( mFDR(\delta^T) \leq \alpha \). \( \square \)

### A.4 Optimality of adaptive procedure

**Lemma 1.** Under the same data generating process and conditions in Theorem 2. For any time stage \( T \), conditioning on the active set \( A_T \in \mathcal{F}_{T-1} \), there exists a \( \lambda^* > 0 \) such that \( \delta^B(\lambda^*, T) \) is the stage-wise optimal test in the sense that it solves the constraint optimization problem

\[
\min_{\{\delta_i\}_{i \in A_T}} \mathbb{E} \left[ \sum_{i \in A_T} \theta_i (1 - \delta_i) \right] \\
\text{s.t. } mFDR(\delta) = \frac{\mathbb{E} \left[ \sum_{i \in A_T} (1 - \theta_i) \delta_i \right]}{\mathbb{E} \left[ \left( \sum_{i \in A_T} \delta_i \right) \lor 1 \right]} \leq \alpha. \tag{14}
\]

Moreover, if we let \( \delta^T = (\delta_1^T, \ldots, \delta_m^T) \) where \( \delta_i^T = 1 \) for \( i \notin A_T \) and \( \delta_{A_c}^T = \delta^B(\lambda^*, T) \), then we also have global \( mFDR \) control at stage \( T \), i.e. \( mFDR(\delta^T) \leq \alpha \).

**Proof.** Given an active set \( A_T \), the coordinates for test at stage \( T \) are fixed and are sampled at all past time stages. Therefore by Theorem 2 with \( m = |A_T| \), we have directly the first part of the constraint optimization problem.

For global \( mFDR \) control, notice that

\[
mFDR(\delta^T) = \frac{\mathbb{E} \left[ \sum_{i=1}^m (1 - \theta_i) \delta_i^T \right]}{\mathbb{E} \left[ \left( \sum_{i=1}^m \delta_i^T \right) \lor 1 \right]} = \frac{\mathbb{E} \left[ \sum_{t=1}^T \sum_{i \in A_t} (1 - \theta_i) \delta_i^T \right]}{\mathbb{E} \left[ \sum_{t=1}^T \left( \sum_{i \in A_t} \delta_i \right) \lor 1 \right]}.
\]

Given stagewise \( mFDR \) control \( \mathbb{E} \left[ \sum_{i \in A_t} (1 - \theta_i) \delta_i^T \right] \leq \alpha \mathbb{E} \left[ \sum_{i \in A_t} \delta_i^T \right] \), we have global \( mFDR \) control. Since the choice of each \( \lambda^* \) in all past time stages controls stagewise \( mFDR \), we have global \( mFDR \) control. \( \square \)

### A.5 Proof of Theorem 3

**Proof.** To show validity of AMSET for \( mFDR \) control at any stopping time \( \tau \), we borrow a similar idea from Wang and Sun (2017). Let \( x^{1:T} = \{x_{it}\}_{i \in A_t, t \leq T} \) be all observations in past stages before time \( T \) and \( \mathcal{F}^T = \sigma(x^{1:T}) \). Recall the definition of \( mFDR \) in the sequential context for stopping time \( \tau \).

\[
mFDR(\tau) = \frac{\mathbb{E} \left[ \sum_{i=1}^\tau \sum_{i \in S_t} (1 - \theta_i) \delta_i^T \right]}{\mathbb{E} \left[ \left( \sum_{t=1}^\tau \sum_{i \in S_t} \delta_i^T \right) \lor 1 \right]}.
\]
Note that the numerator can be written as,

\[
\mathbb{E} \left[ \sum_{t} \sum_{i \in S_t} (1 - \theta_i) \delta_t^i \right] = \mathbb{E} \left[ \sum_{t=1}^{\infty} \sum_{i \in S_t} (1 - \theta_i) \delta_t^i 1\{t \leq \tau\} \right] \\
= \sum_{t=1}^{\infty} \mathbb{E} \left[ \sum_{i \in S_t} (1 - \theta_i) \delta_t^i 1\{t \leq \tau\} \right] \\
\overset{(1)}{=} \sum_{t=1}^{\infty} \mathbb{E} \left[ \sum_{i \in S_t} (1 - \theta_i) \delta_t^i 1\{t \leq \tau\} \mid \mathcal{F}^t \right] \\
\overset{(2)}{=} \sum_{t=1}^{\infty} \mathbb{E} \left[ 1\{t \leq \tau\} \mathbb{E} \left[ \sum_{i \in S_t} (1 - \theta_i) \delta_t^i \mid \mathcal{F}^t \right] \right] \\
\overset{(3)}{=} \sum_{t=1}^{\infty} \mathbb{E} \left[ 1\{t \leq \tau\} \sum_{i \in S_t} \delta_t^i \mathbb{E} \left[ (1 - \theta_i) \mid \mathcal{F}^t \right] \right] \\
\overset{(4)}{=} \sum_{t=1}^{\infty} \mathbb{E} \left[ 1\{t \leq \tau\} \sum_{i \in S_t} \delta_t^i T_{OR}^t \right] \\
\overset{(5)}{\leq} \alpha \sum_{t=1}^{\infty} \mathbb{E} \left[ 1\{t \leq \tau\} \sum_{i \in S_t} \delta_t^i \right] = \alpha \mathbb{E} \left[ \sum_{t=1}^{\tau} \sum_{i \in S_t} \delta_t^i \right] ,
\]

where (1) is by tower property, (2) and (3) are by the fact that both \(1\{t \leq \tau\}\) and \(\delta_t^i\) are \(\mathcal{F}^t\)-measurable, (4) is by definition of \(\text{Ifdr} T_{OR}^t\) in Eq. (5), and (5) is by compound thresholding property in Algorithm 1. Combining the upper bound with the denominator of mFDR, we have that \(\text{mFDR} \leq \alpha\) for any stopping time \(\tau\). This result ensures validity up to continuous monitoring and therefore immunity to peeking.

Stagewise mFDR control is immediate by the same proof as above without summation over time stages. For stagewise exact FDR control, we have that for any time stage \(t\)

\[
\text{FDR}(\delta^i \mid S_t) = \mathbb{E} \left[ \frac{\sum_{i \in S_t} (1 - \theta_i) \delta_t^i}{\left( \sum_{i \in S_t} \delta_t^i \right) \lor 1} \right] \\
= \mathbb{E} \left[ \sum_{i \in S_t} \mathbb{E} \left[ \frac{(1 - \theta_i) \delta_t^i}{\left( \sum_{i \in S_t} \delta_t^i \right) \lor 1} \mid \mathcal{F}^t \right] \right] \\
= \mathbb{E} \left[ \sum_{i \in S_t} \frac{\delta_t^i T_{OR}^t}{\left( \sum_{i \in S_t} \delta_t^i \right) \lor 1} \right] \leq \alpha \mathbb{E} \left[ \frac{\sum_{i \in S_t} \delta_t^i}{\left( \sum_{i \in S_t} \delta_t^i \right) \lor 1} \right] \leq \alpha.
\]

\[\square\]

B Additional procedures

We provide details of some additional procedures in the paper, including non-adaptive version of AMSET called multistage empirical Bayes test (MSET), and Gaussian mixture deconvolution for
estimation of alternative density.

B.1 MSET

The non-adaptive MSET procedure is described in Algorithm ??.

Algorithm 3: MSET

Specify any stopping criterion;

for time stages \( t = 1, 2, \ldots \) do

(Data Sampling) Observe data \( \{x_{it}\}_{i \in [m]} \);

(Ranking) For all \( i \), compute \( T_{OR}^{(i),t} \) in Eq (5) based on \( \{x_{is}\}_{s \leq t} \) and sort them in ascending order

\[
T_{OR}^{(1),t} \leq \cdots \leq T_{OR}^{(m),t}
\]

(Compound Thresholding) Define threshold

\[
T_t = \max \left\{ r : \frac{1}{r} \sum_{i=1}^{r} T_{OR}^{(i),t} \leq \alpha \right\}
\]

Let \( S_t = \{ i \in [m] : T_{OR}^{(i),t} \leq T_{OR}^{(r_t),t} \} \);

(Decision) Let \( \delta_t^i = 1 \) for all \( i \in S_t \) and 0 otherwise;

Break if stopping criterion is satisfied.

B.2 Gaussian mixture deconvolution

In Section 5, we described a data-driven method for AMSET, where we replace the oracle lfdr test statistic in Eq. (5) with estimated population level parameters to form the data-driven test statistic in Eq. (11). In particular for the estimation of non-null density, we will use Kiefer-Wolfowitz maximum likelihood estimation for Gaussian mixtures with software “REBayes” developed by Koenker and Gu (2017) and computation supported by Koenker and Mizera (2014).

Recall that Gaussian mixture deconvolution gives us marginal density estimation as a mixture of Gaussians \( \hat{G} = \sum_{i=1}^{k} \hat{p}_i \hat{G}_i \), where \( \sum_{i=1}^{k} \hat{p}_i = 1 \) and \( \hat{G}_i = \mathcal{N}(\hat{\mu}_i, 1) \) for some \( k \). There are various methods for recovering an estimate of \( f_1 \) from deconvolution results. We list three possible methods below.

- Impose a hard threshold \( c \) such that we only include \( \hat{G}_i \)'s such that \( |\hat{\mu}_i| \geq c \) with their weighted corresponding proportions.

- Use the classification boundary in homoscedastic Gaussian mixture model in Donoho and Jin (2004) as a sanity check for estimation of densities. That is, given an estimate of the proportion of non-nulls, we can find the smallest detectable effect from the classification boundary and consider only detectable effects as potential candidates for alternative density estimation.

- Estimate threshold to recover alternative density in a completely data-driven way. That is, we first sort components of the Gaussian mixture estimate in ascending order of \( |\hat{\mu}_i| \)'s, resulting
in $\hat{G}(1), \ldots, \hat{G}(k)$ with their corresponding proportions $\hat{p}(1), \ldots, \hat{p}(k)$. We can place a threshold $k_a = \max\{k : \sum_{i=1}^{k} \hat{p}(i) \leq 1 - \hat{p}\}$, where $\hat{p}$ is the proportion of non-nulls estimated by step (1) of data-driven implementation in Section 5. The alternative density can then be estimated by $\hat{f}_1 = \sum_{i=k_a}^{k} \hat{p}(i) \hat{G}(i) / \sum_{i=k_a}^{k} \hat{p}(i)$.

C Reproducibility

We provide R implementations of all models described and the reproduction of all simulation study results at https://github.com/AMSET2022/amset. This supplementary section details the steps to reproduce our results.

C.1 Code Organization

All R codes for generating simulation results are included in the folder “R”. The csv files and plots for our simulations in fixed horizon comparison and stagewise comparison can be found in the “data” folder. Since the package REBayes that provides Kiefer-Wolfowitz maximum likelihood estimation for mixture models deconvolution requires the installation of the optimization package Rmosek. We also provided links to user guide and resources for installing Rmosek.

C.2 Simulation code

The details for simulation code are as follows.

- **epsest.func.R** contains auxiliary function to estimate the proportion of non-null locations.
- **MSET.R** implements fixed horizon comparison between MSET procedure and Optimizely’s method using always valid p-values.
- **MSET_stagewise.R** implements stagewise comparison between MSET procedure and Optimizely’s method using always valid p-values.
- **AMSET.R** implements fixed horizon comparison between AMSET procedure and Optimizely’s method using always valid p-values.