Sequential Multiple Hypothesis Testing with Type I Error Control

Alan Malek
MIT
amalek@mit.edu

Sumeet Katariya
University of Wisconsin-Madison
katariya@wisc.edu

Yinlam Chow
Stanford University
ychow@stanford.edu

Mohammad Ghavamzadeh
Adobe Research
ghavamza@adobe.com

Abstract

This work studies multiple hypothesis testing in the setting when we obtain data sequentially and may choose when to stop sampling. We summarize the notion of a sequential p-value (one that can be continually updated and still maintain a type I error guarantee) and provide several examples from the literature. This tool allows us to convert fixed-horizon step-up or step-down multiple hypothesis testing procedures (which includes Benjamini-Hochberg, Holm, and Bonferroni) into a sequential version that allows the statistician to reject a hypothesis as soon as the sequential p-value reaches a threshold while maintaining type I error control. We show that if the original procedure has a type I error guarantee in a certain family (including FDR and FWER), then the sequential conversion inherits an analogous guarantee. The conversion also allows for allocating samples in a data-dependent way, and we provide simulated experiments demonstrating an increased number of rejections when compared to the fixed-horizon setting.

1 Introduction

Hypothesis testing (HT) is a statistically rigorous procedure that makes decision about the truth of a hypothesis is a way that allows the probability of error to be carefully controlled. Specifically, given a null hypothesis and an alternative hypothesis, the statistician decides when to reject the null hypothesis such that the type I error, the probability of falsely rejecting the null hypothesis, is bounded. A second goal is to minimize the probability type II error, which occurs when the statistician erroneously fails to reject the null (when the alternative is indeed true). HT is a fundamental problem in statistical inference that arises in numerous fields. Examples include online marketing [6] and pharmaceutical studies [20].

HT has been traditionally studied in the *fixed-horizon* setting. The desired parameters such as type I error and power are used to calculate a lower bound on the necessary sample size, the requisite number of samples are collected, and then a decision is made. In particular, one must wait until the minimum sample size is reached or the type I error guarantee is forgone. While this approach provides rigorous statistical guarantees in accuracy and robustness, it usually takes a long time to conclude a test and also its offline nature possesses huge limitations to emerging applications, such as digital marketing, where data is streaming in a sequential fashion and a data-efficient approach is crucial to allow real-time adaption to user's preferences.

Multiple hypothesis testing (MHT) is a natural extension of HT to control error across multiple hypotheses. While type I and type II errors are straightforward in single tests, there are many ways one could penalize incorrect decisions in MHT. The two most widely used notions are family-wise error rate (FWER), which is the probability of at least one false rejection, and false discovery rate (FDR), which is the expected proportion of the rejections that are false. Generally, the statistician would like to reject as many tests as possible while still controlling the number of false positives.

Recent advances in the scale and rate of data collection have increased the need for *sequential* hypothesis testing; in this framework, the data arrives sequentially and the statistician may make decisions after observing partial data, such as deciding when to stop collecting data or from which populations to collect data from.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

Large clinical trials and Hypothesis testing in social networks, mobile applications, and large internet systems are examples where sequential HT is desirable. In contrast, in fixed-horizon setting all the data are collected a priori.

Of course, such data-dependent decision introduce dependencies between future data and past data, and the statistician must be more careful designing procedures if type I error control is desired. For example, a common goal in sequential hypothesis testing is to observe the data sequence and stop at the statisticians choosing. However, if the statistician were to follow the evolution of a statistic in the traditional fixed-horizon HT and reject the null hypothesis as soon as it crosses the fixed-horizon decision boundary, the type I error would be greatly inflated [18] and many erroneous rejections would be made. We survey some alternative hypothesis testing procedures that provide a type I error guarantee in the sequential setting, but our work will use the existence of such sequential tests as a starting point. Our goal will be to use sequential hypothesis tests to adapt multiple hypothesis testing procedures into sequential variants that can enjoy the benefits of data-depending sampling and early stopping.

1.1 Related Work

Sequential hypothesis testing has a long history [30, 20, 24], but its extension to multiple hypotheses and tests is more recent. Several recent papers study the mechanics of statistics that allow for uniform type I error guarantees for any stopping time [1, 15, 28, 29].

Bartroff and co-authors have established several scenarios where sequential MHT has type I error in the cases of all simple hypotheses with FWER [2] and FDR control [3]. Similarly, De and Baron [10] proposed a sequential MHT algorithm that guarantees the overall type I error. However since this algorithm is based on the Bonferroni correction [9], the power of rejecting alternative hypotheses has been empirically shown to be low. Fellouris and Tartakovsky [14] also studied the sequential testing problem of a simple null hypothesis vs. a composite alternative hypothesis, for which the decision boundaries were derived using both mixture-based and weighted-generalized likelihood ratio statistics. While they have shown asymptotic optimally of such a sequential testing procedure, their proposed method requires high-dimensional hyper-parameter tuning and cannot be easily extended to multiple hypothesis testing. Zehetmayer and collaborators [31, 32] looked at controlling FDR when the p-values are modeled as Gaussian.

Similar to our results, the approach of [18] relies on sequential p-values. Our work extends their results by allowing each test to be stopped individually. This dynamic allocation property allows us to conclude more tests (than their static/uniform allocation scenario) at each time-step during the lifespan of the procedure. Moreover, our framework is more flexible and allows control of FDR, FWER, and a host of other metrics, in contrast to only controlling FDR.

1.2 Our Contributions

We first describe a framework that fits most fixed-horizon hypothesis test procedures, including all *step-up* procedures (e.g., Hochberg [16] and Benjamini-Hochberg [7]) and *step-down* procedures (e.g., Holm's [17]) with an accompanying family of type I errors that includes FWER and FDR. By leveraging the powerful tool of sequential p-values, we propose a method for converting any fixed-horizon MHT into a sequential MHT and show that the fixed-horizon guarantees generalizes to the sequential setting in Section 5.2. Further, our sequential MHT can stop tests early in a data-dependent manner, and therefore stops easier hypotheses first and potentially allowing more samples for difficult hypotheses. We discuss this dynamic allocation property and demonstrate that it does lead to more rejected tests on synthetic data in Section 7. Additionally, Section 6 discusses controlling type II errors and provides a sequential calculator for estimating a minimum sample size to control both FDR and false non-discovery rate (FNR) in sequential MHT.

2 Hypothesis Testing

In hypothesis testing (HT), the statistician tests a *null hypothesis* H_0 against an *alternative hypothesis* H_1. The goal is to design a procedure that controls the type I error α, the probability of erroneously rejecting H_0, while simultaneously minimizing the type II error β, the probability of failing to reject H_0 when it is false. A prototypical example is A/B testing with baseline option A and alternative option B; the null and alternative hypotheses are $H_0 : \mu_A = \mu_B$ and $H_1 : \mu_A \neq \mu_B$, where μ_A and μ_B are the mean values (e.g., conversion rates or click through rates).

Multiple hypothesis testing (MHT) is an extension of HT to a finite number of tests m. Examples are a base option A against a set of m alternative options $A_1, \ldots, A_m$, a single null hypothesis against a set of alternative hypotheses [23, 5], and a test with a set of nested hypotheses [19]. When the test stops, the MHT procedure rejects R tests, out of which V are rejected by mistake, and does not reject $m - R$ tests, out of which W should have been rejected. Table 1 summarizes these quantities. The decision of the MHT procedure to reject/not-reject a test is a random quantity, because it depends on the test sample that is random. Thus, all

	not-rejected	rejected	total
H_0 true	U	V	m_0
H_0 false	W	S	$m - m_0$
total	$m - R$	R	m

Table 1: MHT Error Quantities.

the elements in Table 1, except m and m_0 are random variables.

Unlike HT, there is no single notion of type I error for multiple hypotheses. The two common are family-wised error rate (FWER) and false discovery rate (FDR) that are defined as[1]

$$\mathsf{FWER} := P(V \geq 1) \quad \text{and} \quad \mathsf{FDR} := \mathbb{E}\left[\frac{V}{R \vee 1}\right]. \quad (1)$$

The goal is usually to control these metrics at the level of q. FWER, first introduced by Bonferroni [9], is the probability of making at least one false positive. As the number of tests grows, requiring FWER control quickly makes rejection very difficult. Therefore, FDR [7], which controls the expected proportion of false discoveries, has become a popular alternative as it allows the number of false positives to grow with the number of tests. This notion has been expanded to dependent tests [8] and to variations such as γ-FDR [21].

One can also ask for guarantees on the type II error (see e.g., [26, 27]) by defining

$$\mathsf{FWER\ II} := P(W \geq 1) \quad \text{and} \quad \mathsf{FNR} := \mathbb{E}\left[\frac{W}{(m-R) \vee 1}\right].$$

We do not consider procedures with type II error bounds in this paper, as they require strong assumptions on the distributions of the statistics under the alternative hypothesis.

2.1 Fixed-Horizon Hypothesis Testing

Fixed-horizon is the traditional approach to HT that is widely used in industry. In the fixed-horizon HT, the statistician calculates the minimum necessary sample size for some desired type I and type II errors (plus some information that depends on the hypotheses being tested). The test should be continued until this horizon is reached. At this point, a typical fixed-horizon HT procedure first computes the test statistic from the observations and then uses it to calculate the p-value of the test, i.e., the probability under the null hypothesis of sampling a test statistic at least as extreme as the one that was observed. Finally, it rejects the null hypothesis if the p-value is less than or equal to the desired type I error, i.e. $p \leq \alpha$.

[1]We denote by $\vee$ the maximum operator, e.g., $R \vee 1 = \max\{R, 1\}$.

The most common procedure for **fixed-horizon multiple hypothesis testing** is to pick sample sizes, gather the samples and compute the p-values of the m tests, then determine, as a whole, which null hypotheses to reject. The most common MHT procedure to control FWER is Bonferroni [9], which controls FWER at the level q by applying the union bound and only rejecting the tests whose p-values are smaller than q/m. However, Bonferroni procedure lacks power and is overly conservative, especially when the number of hypotheses is large or the test statistics are highly correlated [22]. Popular alternatives to Bonferroni to control FWER are Holm's *step-down* procedure [17] that has larger power than Bonferroni with the same type I guarantee, and Hochberg's *step-up* procedure [16] that rejects even more tests than Holm's with the same guarantees. FDR is typically controlled by Benjamini-Hochberg's procedure [7], which is *step-up*. Formally,

Definition 2.1 (MHT Procedures)**.** *A MHT procedure $\mathcal{P}$ consisting of m tests is a mapping $[0,1]^m \to \{0,1\}^m$ from the set of p-values of the individual tests to m reject/not-reject decisions.*

*Denote the ascending p-values of the tests by $p^{(1)} \leq \ldots \leq p^{(m)}$. We say that $\mathcal{P}$ is a **step-up** procedure, if for some sequence of decision thresholds $\alpha_1, \ldots, \alpha_m$, $\mathcal{P}$ rejects tests $(1), \ldots, (k^*)$ for*

$$k^* = \max\{k : p^{(k)} \leq \alpha_k\}.$$

*Similarly, $\mathcal{P}$ is a **step-down** procedure, if for some sequence of decision thresholds $\alpha_1, \ldots, \alpha_m$, $\mathcal{P}$ rejects tests $(1), \ldots, (k^* - 1)$ for*

$$k^* = \min\{k : p^{(k)} > \alpha_k\}.$$

*Finally, we will call $\mathcal{P}$ a **monotonic** test procedure, if it is either step-up or step-down.*

The step-up procedures Hochberg and Benjamini-Hochberg set their decision thresholds to $\alpha_k = \frac{q}{m-k}$ and $\alpha_k = \frac{kq}{m}$, respectively, and the step-down procedure Holm sets its decision threshold to $\alpha_k = \frac{q}{m-k}$. Note that q is the desired FWER level in Hochberg and Holm procedures, and the desired FDR level in Benjamini-Hochberg.

The following definition encapsulate a large set of MHT type I metrics for different functions f. This unified formulation will help us with our analysis in Section 5.2.

Definition 2.2 (Error Guarantee)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}^+$ be a monotonically increasing function in its first argument and the error level $q > 0$ be a positive real number. We say that a test procedure $\mathcal{P}$ has an (f, q) error guarantee if*

$$\mathbb{E}\left[f(V, S)\right] \leq q \quad (2)$$

for all values of m and m_0, and all distributions of true rejections S, assuming that the distributions of the p-values corresponding to the m_0 true null hypotheses $p^{\pi(1)}, \ldots, p^{\pi(m_0)}$ are marginally uniform on $[0,1]$, where π is any arbitrary permutation of $1, \ldots, m_0$.

If Eq. 2 holds only when $p^{\pi(1)}, \ldots, p^{\pi(m_0)}$ are i.i.d. uniform, we will say that $\mathcal{P}$ has an (f,q) error guarantee under independence (a weaker condition).

For the well-known MHT type I error metrics FWER and FDR, function f is equal to $f(V,S) = \mathbf{1}[V > 0]$ and $f(V,S) = \frac{V}{(V+S) \vee 1}$, respectively.

3 Problem Formulation

Consider testing the null hypothesis H_0 against the alternative hypothesis H_1 with type I error α. Let $X \in \mathcal{X}$ be the random variable of test samples and x be a realization of this random variable. We start by defining the notion of sequential p-value, the crucial tool that we make use of throughout the paper.

Definition 3.1. *A sequential p-value for testing the null hypothesis H_0 against the alternative hypothesis H_1, denoted by $\{p_t\}$, is a sequence of mappings $p_t : \mathcal{X}^t \mapsto [0,1]$, $t \geq 1$ that satisfy the following two properties* ***under the null hypothesis***:[2]

1. *(super-uniform) For any $\delta \in [0,1]$ and any $t \geq 1$, we have*

$$P\left(\sup_{s \leq t} p_s(X_1, \ldots, X_s) \leq \delta\right) \leq \delta. \tag{3}$$

2. *(non-increasing) For any fixed realization $\{x_t\}_{t \geq 1} \in \mathcal{X}^\infty$ and any $t \geq 1$, we have*

$$p_t(x_1, \ldots, x_t) \geq p_{t+1}(x_1, \ldots, x_{t+1}).$$

Sequential p-value allows "peeking" into the test, i.e. rejecting the null hypothesis without violating the type I error guarantee. If we reject H_0 as soon as $p_t \leq \alpha$, the super-uniform property guarantees that the type I error remains below α. Note that unlike the sequential setting, peeking in fixed-horizon hypothesis testing can greatly increase the type I error. This is due to the fact that in this class of tests, the p-value has been designed to hold only at the test horizon and not uniformly across the entire sample path.

Our framework is agnostic to the hypothesis test as long as we have access to a sequential p-value for the test. There is an extensive literature on creating sequential p-values for hypothesis tests. While this is not the focus of the paper, we present a few examples in Section 4 that include Wald's sequential probability ratio test, and Bayes factor under a simple null hypothesis.

[2]Note that a sequential p-value is a sequence of random variables whose randomness comes from the samples X.

Problem definition: We can now state the problem that we study. Suppose that the statistician has a family of m simultaneous sequential hypothesis tests and her goal is to establish a procedure for rejecting tests while controlling the type I error across the whole family. For the hypothesis test k, we denote by H_0^k the null hypothesis, H_1^k the alternative hypothesis, and $\{p_t^k\}$ the sequential p-value. At each time t during the test, the statistician has access to the sequential p-values of all the m tests, $\{p_t^1\}, \ldots, \{p_t^m\}$, and can generate new samples and update p-values arbitrarily. Her goal is to select samples in order to maximize the number of rejected null hypotheses without violating the considered notion of type I error over the family of tests (e.g., FWER or FDR).

4 Examples of Sequential P-value

In the previous section, we defined sequential p-values without arguing about their existence. This section describes several sequential p-values and provides the reader with more references. We also prove some important properties.

4.1 Sequential Probability Ratio Tests

The sequential probability ratio tests (SPRT) [30] was one of the first sequential hypothesis testing frameworks where only two hypotheses are taken into the account. In order to extend the aforementioned SPRT techniques to address the sequential multiple hypothesis testing problem, Robbins and Siegmund [25] developed the $M-$ary sequential probability ratio test (MSPRT) procedure that leverages the notion of mixture posterior probabilities in the likelihood ratio test. In particular, the stopping rule effectively picks one of the M hypotheses that has the largest posterior probability, given all the previous samples observed. More recent work on the extension of MSPRT can also be found in [5, 12, 11].

Recall the infinite sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$ with an underlying probability density function f. Also let H_j be the hypothesis that $f = f_j$, for $j \in \{1, \ldots, M\}$, such that $f_k \neq f_j$, almost surely for all $j \neq k$. We further assume that the prior probabilities of the hypotheses are known, and let π_j denote the prior probability of hypothesis H_j for each j. Therefore, given a sequence of thresholds $\{A_1, \ldots, A_M\}$ that characterizes the false discovery rate of each hypothesis, the stopping time T for the MSPRT can be described

as the minimum sample $n \geq 1$ such that

$$\frac{\pi_k \prod_{i=1}^n f_k(X_i)}{\sum_{j=1}^M \pi_j \prod_{i=1}^n f_j(X_i)} > \frac{1}{1+A_k}.$$

for some $k \in \{1, \ldots, M\}$. Correspondingly, the final decision is defined as $\delta = H_k$.

For comparison, Let A_0 and $A_1 > 0$ be the specific parameters that characterize the type-I and type-II error guarantees of a sequential two hypothesis testing procedure, we have the SPRT [30] stopping time T defined as the minimum sample $n \geq 1$ such that

$$L_n = \frac{\prod_{i=1}^n f_1(X_i)}{\prod_{i=1}^n f_0(X_i)} \notin [A_0, A_1].$$

Here the final decision is given by $\delta = H_1$ if $L_n > A_1$, and it is given by $\delta = H_0$ if $L_n < A_0$. It is straightforward to show that the MSPRT with $M = 2$ is identical to the SPRT with parameters $\pi_0 A_0/\pi_1$ and $\pi_0/(\pi_1 A_1)$. This implies that for the SPRT, the prior probabilities can be incorporated into the parameters A_0 and B'.

4.2 Test Martingales

The construction of p–tests is an important problem but not the focus of this paper. We refer the reader to [29, 15] for a more thorough treatment but will include a few examples below. We are particularly interested in test super-martingales, which are nonnegative super-martingales $\boldsymbol{X}_t \geq 0$ for any t that satisfy $\mathbb{E}[\boldsymbol{X}_0] \leq 1$, and test martingales, which are martingales that are nonnegative and satisfy $\mathbb{E}[\boldsymbol{X}_0] = 1$. We require test martingales have initial value 1. A simple application of the optional stopping theorem yields the following well known result, often called the maximal inequality. See Appendix A for a proof.

Lemma 4.1. *For any test martingale $\boldsymbol{X}_t$, we have $P\left(\sup_t \boldsymbol{X}_t > b\right) \leq \frac{1}{b}$.*

This inequality shows that $\boldsymbol{X}_t$ can take the value ∞ only with probability zero. Now we will connect the notion of test super-martingale and sequential p-value by showing that the inverse of a supremum of a test super-martingale is indeed a sequential $p-$value. This statement is true when the supremum is taken over all time points, and its proof can be found in Appendix A.

Corollary 4.2. *If the test-statistic Λ_t is a positive test super-martingale, then $\frac{1}{\sup_{n' \leq t} \Lambda_{n'}}$ is a sequential p-value.*

While the above result shows that a sequential $p-$value can be easily derived from a test super-martingale, in many likelihood ratio based sequential hypothesis testing (i.e. SPRT) procedures, finding a test super-martingale is still a non-trivial task. In order to have a more natural construction of the sequential $p-$value in many of such tests, we will introduce the concept of Bayes factor and draw some important connections between Bayes factor, test super-martingale and the sequential $p-$value. Consider an arbitrary probability space $(\Omega, \mathcal{F}, P)$, a non-negative measurable function $B : \Omega \to [0, \infty]$ is known as a Bayes factor for probability measure P if $\int (1/B) dP \leq 1$. A Bayes factor B is said to be precise if $\int (1/B) dP = 1$. In order to relate this definition of Bayes factor with the traditional definition of a likelihood ratio, we note first that whenever Q is a probability measure on $(\Omega, \mathcal{F})$, the Radon-Nikodym derivative (or the likelihood ratio) dQ/dP will satisfy $(dQ/dP)dP \leq 1$, with equality if Q is absolutely continuous with respect to P. Therefore, by definition $B = 1/(dQ/dP)$ is a Bayes factor for P. The Bayes factor B will be precise if Q is absolutely continuous with respect to P. In this case B will be a version of the Radon-Nikodym derivative dP/dQ.

Equipped with the mathematical definition of a Bayes factor, we now provide the following result that shows for a sequential hypothesis test (A/B test) with a simple null-hypothesis, the Bayes factor defined under the null hypothesis is a test martingale.

Corollary 4.3. *Consider any sequential hypothesis tests with an alternative hypothesis H_1 and a simple null hypothesis H_0. Both the likelihood ratio under $H_0 : \theta_0$ and Bayes factor under $H_0 : \theta_0$ are test martingales.*

The proof of this corollary can be found in Appendix A. Combining this result with the aforementioned one from Corollary 4.2, one concludes that the reciprocal of the supremum of Bayes factor (or likelihood ratio) over all time points is a sequential $p-$value.

4.3 Empirical Approaches

Besides using MSPRT or test martingales to construct decision boundaries for sequential hypothesis testing, one can also construct non-parametric methods for sequential A/B testing [1]. By keeping track of a single scalar test statistic that is derived based on a zero-mean random walk process under the null hypothesis, this non-parametric procedure tests the null hypothesis whenever a new data point is processed, and once a hypothesis is rejected it controls the type I error by utilizing the classical probability result of the law of the iterated logarithm (LIL). Since this procedure is sequential, by nature it only takes linear time and constant space to compute the decision at each step. Furthermore, by using the empirical Bernstein-LIL-based analysis from the above paper, it has also been shown that this algorithm has the same power guarantee as

its non-sequential counterpart.

5 Algorithms for Sequential Multiple Hypothesis Testing

Now that we have presented a framework for fixed-horizon MHT and sequential p-values, we can describe our procedure for conducting MHT in the sequential setting. Here, we assume that, for each hypothesis k, the decision maker has access to a sequential p-value $p_1^k, p_2^k, \ldots$. In the fixed horizon MHT, the statistician collects samples from each hypothesis, calculated the corresponding p-value, then applies a testing procedure; we would like to applying a rejection procedure before all the samples are collected with the hope that easier tests can be concluded early. The intuition is that, because sequential p-values p_t^k are valid for all t and non-decreasing, we may apply any monotonic rejection procedure iteratively without distorting which tests are rejected. Hence, we need only sample from the tests that have not been rejected. This section is devoted to proving that this intuition does hold, and indeed the sequential procedure inherits the same (f, q) error guarantee as the fixed-horizon procedure.

Specifically, we begin by defining the sequential analog of a test procedure and then show how sequential p-values can be used to upgrade a fixed horizon procedure with a (f, q)-guarantee into a sequential test procedure with an analogous guarantee. This result is the main theorem of this paper and is presented in Section 5.2.

Definition 5.1 (Sequential Test Procedure)**.** *Given m sequential p-values $\{p_t^1, \ldots, p_t^m\}_{t\geq 1}$, a sequential test procedure consists of, for every round $t = 1, 2, \ldots$,*

- *Samplers $\mathcal{S}_t : \{p_s^k\}_{s\leq t-1, 1\leq k\leq m} \mapsto \{0,1\}^m$ to decide which hypothesis tests to sample from during round t,*
- *A stopping time T with respect to the filtration generated by $\{p_t^1, \ldots, p_t^m\}$ (i.e. T is measurable w.r.t. this filtration) that determines when to stop sampling, and*
- *A mapping $\mathcal{P} : [0,1]^{m\times\infty} \mapsto \{0,1\}^m$ from histories of sequential p-values to reject/not-reject decisions.*

In words, the sequential test procedure $(\mathcal{S}_t, T, \mathcal{P})$ dictates what hypotheses to sample from, when to stop, and what hypotheses to reject once stopped.

Further, the analogous quantities to U, V, and R in Figure 1 will be denoted as U_T, V_T, and R_T to emphasize that these are the random variables evaluated when the testing procedure finishes at stopping time T.

Fixed-horizon The fixed-horizon multiple hypothesis test with horizon N corresponds to the sequential procedure with $p_t^k = p_N^k$ (the p-values are constant), $\mathcal{S}_t = 1^m$ (every test is sampled every round), and $T = N$ (the stopping time is deterministic).

Our definition of sequential test procedure provides an easy extension of the (f, q) guarantee for fixed-horizon:

Definition 5.2. *Given a function $f : \mathbb{R}^2 \to \mathbb{R}^+$ that is monotonically increasing in the first argument and a real number $q > 0$, we say that a sequential test procedure $(\mathcal{S}_t, T, \mathcal{P})$ has an (f, q) error guarantee if*

$$\mathbb{E}[f(V_T, S_T)] \leq q \tag{4}$$

for all m_0 and all distributions of true rejections S, assuming that p_t^k is a sequential p-values for every true null hypothesis.

If $\mathbb{E}[f(V_T, S_T)] \leq q$ only holds when the sequential p-values corresponding to the true null hypothesis are independent, we will say that $(\mathcal{S}_t, T, \mathcal{P})$ has a error guarantee under independence (a weaker condition).

5.1 The Sequential Conversion

We now have an idea of what a sequential test procedure consists of and what type of error guarantee we can hope to achieve. This section describes a method to transform a fixed-horizon test procedure into a sequential test procedure, and the next section shows that this sequential conversion method transforms fixed-horizon procedures with (f, q)-guarantees into sequential procedures with (f_T, q)-guarantees.

Let $\mathcal{F}_t$ be the filtration generated by $\{p_t^1, \ldots, p_t^m\}$ and consider a test procedure $\mathcal{P}$. By a slight abuse of notation, we say that $i \in B$, where $B \in \{0,1\}^m$ is some binary vector, if $B_i = 1$. The sequential conversion $\mathcal{C}$ of $\mathcal{P}$ is defined as follows.

Definition 5.3. *Assume we have sequential p-values $p_t^1, \ldots, p_t^m$, and some stopping time T with respect to $\mathcal{F}_t$ with finite expectation. The* sequential conversion *of the test procedure $\mathcal{P}$, which we label $\mathcal{C} = (\mathcal{S}'_t, T, \mathcal{P}')$, is defined to be*

- *For round t, we sample $\mathcal{S}'_t = \mathcal{P}(p_{t-1}^1, \ldots, p_{t-1}^m)$*
- *If $k \in \mathcal{S}'_t$, sample from the hypothesis and update the sequential p-value; otherwise, $p_t^k = p_{t-1}^k$.*
- *At the end of the test, we reject $\mathcal{P}' = \mathcal{P}(p_T^1, \ldots, p_T^m)$.*

Intuitively, $\mathcal{C}$ applies the fixed-horizon procedure $\mathcal{P}$ at every round to the sequential p-values and does not sample from the already rejected hypotheses. The pseudocode for $\mathcal{C}$ is presented in Algorithm 1.

Without the use of sequential p-values, using $\mathcal{C}$ to conduct a hypothesis test would greatly inflated type I errors for the same reasons that peeking invalidates type I error guarantees in single hypothesis testing. However, if $\mathcal{P}$ is of a certain form and we use sequential p-values, we maintain error control in a strong sense: the (f, q) guarantee of $\mathcal{P}$ translates exactly into an (f, q) guarantee of $\mathcal{C}$. This results is precisely stated and proven in the next section. We would like to emphasize that the algorithm allows the user to stop each test once they have reached a significance level while still maintaining a type 1 error guarantee.

Algorithm 1 Sequential Conversion $\mathcal{C}$

```
Input: stopping time T, rejection procedure P, Type 1 error α
Initialize p_0^1, ..., p_0^m to 1
for t = 1, 2, ..., do
    Set S_t = P(p_{t-1}^1, ..., p_{t-1}^m)
    For each k ∉ S_t, draw a sample and update p_t^k
    For each k ∈ S_{t-1}, set p_t^k = p_{t-1}^k
    if Stopping time T is reached or S_t = ∅ then
        break
    end if
end for
Return rejected tests S_{T∨t}
```

5.2 Main Result

Theorem 5.4. *Let $\mathcal{P}$ be a monotonic test procedure with a (f, q) guarantee. Then its sequential conversion $\mathcal{C}$ of $\mathcal{P}$ given by Definition 5.3 also has an (f, q) guarantee. That is,*

$$\mathbb{E}[f(V_T, S_T)] \leq q.$$

Furthermore, if $\mathcal{P}$ only has an independent (f, q) guarantee, then $\mathcal{C}$ only has an independent (f, q)-guarantee.

Before the proof, we give two lemmas. First, we argue that applying a test procedure $\mathcal{P}$ to sequential p–values instead of offline p–values retains the same error guarantee (at the potential cost of lower power). Second, we show that the sequential procedure with early stopping, as described in Algorithm 1, produces the same decisions as the procedure where all sequential statistics are run to the same length. Putting these two facts together, we have that the early stopping sequential procedure inherits the same error guarantee as the test procedure $\mathcal{P}$. Proofs for both lemmas are in the appendix.

Lemma 5.5. *Consider a monotonic rejection procedure $\mathcal{P}$ with an (f, q) guarantee, sequential p–values $p_t^1, \ldots, p_t^m$, and a stopping time T with respect to $\mathcal{F}_t$. Let $\mathcal{R} = \mathcal{P}(p_T^1, \ldots, p_T^m)$, i.e. the results of applying $\mathcal{P}$ to the p–tests at time T. Denoting $V_T = |\{R^k = 1 : k \in H_0\}|$ and $S_T = m_0 - |\{R^k = 1 : k \in H_0\}|$ (the number of false positives and true positives, respectively), we have*

$$\mathbb{E}[f(V_T, S_T)] \leq q. \tag{5}$$

Furthermore, if $\mathcal{P}$ has only an independent (f, q) guarantee, then (5) holds when the p–tests are independent.

Lemma 5.6. *Under the same setting as Lemma 5.5, consider $\mathcal{C}$, the sequential test procedure defined the sequential conversion in Definition 5.3, and $\mathcal{C}'$, defined to be the sequential test procedure that is identical except with $T = N$. Then, for every realization of $p_t^1, \ldots, p_t^m$, $\mathcal{C}$ and $\mathcal{C}'$ reject the same tests.*

Proof of Theorem 5.4. Consider $\mathcal{C}$ and $\mathcal{C}'$ as in Lemma 5.6; let V_T and S_T be the random variables corresponding to $\mathcal{S}(\mathcal{P})$ and V'_N, S'_N the the random variables corresponding to $\mathcal{S}'(\mathcal{P})$.

From Lemma 5.5, we have that $\mathbb{E}[f(V'_N, S'_N)] \leq q$. On the other hand, Lemma 5.6 implies that the decision of both procedures are exactly the same, and hence $V_T = V'_N$ and $S_T = S'_N$ almost surely. Combining these, we have $\mathbb{E}[f(V_T, S_T)] \leq q$. □

6 Sequential Calculator for FWER

Even for the sequential setting, it is desirable to have an an estimate of the number of samples required. While it is standard to estimate the effective horizon in a fixed-horizon setting [13], effective horizon calculations for sequential tests are underdeveloped. Therefore, we propose a sequential calculator for the general MHT procedure. Consider the case when there are m tests with corresponding sequential p-values given by $p_t^1, \ldots, p_t^m$. Equipped with the following parameters: 1) FWER α, and 2) FWER II β, the problem is to find a deterministic horizon N^* such that both FWER and FWER II are guaranteed, while similar performance guarantees are seen in SPRT-based algorithms such as [4] for simple hypotheses.

6.1 Bonferroni Correction

Recall from the case of an A/B test with statistic Λ_t, where the stopping time T is the random time at which the gap first crosses the decision boundary, i.e., $T = \inf_t\{\Lambda_t \geq 1/\alpha\}$.

We can use the Bonferroni correction to extend this property to multiple hypothesis testing and obtain a stopping condition that guarantees a total FWER of α: $\Lambda_t^k \geq m/\alpha$, $\forall k \in \{1, \ldots, m\}$. Here the term m/α corresponds to the union bound with FWER guarantee, and T^k is the stopping time of test k.

Let $\mathcal{K}$ be set of tests with a true alternative hypothesis. We wish to find the smallest horizon N^* such that $\mathbb{P}(T^k \leq N^*,\ \forall k \in \mathcal{K}|\mathcal{K}) \geq 1-\beta$. Because the underlying hypothesis of each test is generally unknown, we cannot solve for N^* exactly and must choose N^* to hold for all possible values of $\mathcal{K}$. By the union bound for intersection of events, we have that

$$\begin{aligned}\mathbb{P}\left(T^k \leq t, \forall k \in \mathcal{K}|\mathcal{K}\right) &= \mathbb{P}\left(\bigcap_{k\in\mathcal{K}} \{T^k \leq t\}|\mathcal{K}\right) \\ &\geq \sum_{k\in\mathcal{K}} \mathbb{P}(T^k \leq t|\mathcal{K}) - (|\mathcal{K}|-1) \\ &= \sum_{k\in\mathcal{K}} \mathbb{P}(T^k \leq t|H_1^k) - (|\mathcal{K}|-1),\end{aligned}$$

where last equality is due to the independence of tests. When $t = \widetilde{N}^*$, we have that

$$\sum_{k\in\mathcal{K}} \mathbb{P}(T^k \leq \widetilde{N}^*|H_1^k) - (|\mathcal{K}|-1) \geq 1 - |\mathcal{K}|\beta/m \geq 1-\beta,$$

which by definition of N^* implies $\widetilde{N}^* \geq N^*$. Utilizing this property, one can approximate the effective horizon by upper bounding N^* as follows: $\widetilde{N}^* = \max_{k\in\{1,\ldots,m\}} N^{*,k}$ such that

$$N^{*,k} = \inf\left\{t : \mathbb{P}(T^k \leq t|H_1^k) \geq 1 - \frac{\beta}{m}\right\}.$$

7 Experiments

One of the primary advantages of dynamic allocation is that we can stop easy tests early. Consider the scenario when we have a limit on the number of samples but can allocate them to arbitrary hypotheses. In the non-sequential framework, the budget is uniformly allocated to each test, a p-value is computed, and the Benjamini-Hochberg procedure is applied at the end. However, our sequential framework allows sampling to cease for a test that will certainly be rejected at the end of the sequential procedure, a property we refer to as dynamic allocation. Thus, our procedure can use proportionally more samples on harder tests.

We tested 1000 simple vs. composite hypotheses where the null is a zero-mean Gaussian and the alternatives are non-zero mean Gaussians. Of these tests, 800 are true alternatives with means uniformly chosen from $[-10, 10]$ and the other 200 are true nulls (with zero mean). We compare the performance of our sequential algorithm against the classical non-sequential Benjamini-Hochberg procedure [7]. This entire procedure was repeated over 1000 independent instantiations and the results were averaged. The p-values for the non-sequential test were calculated using the independent two-sample t-test for groups with equal variance. The sequential p-values were calculated using the sequential

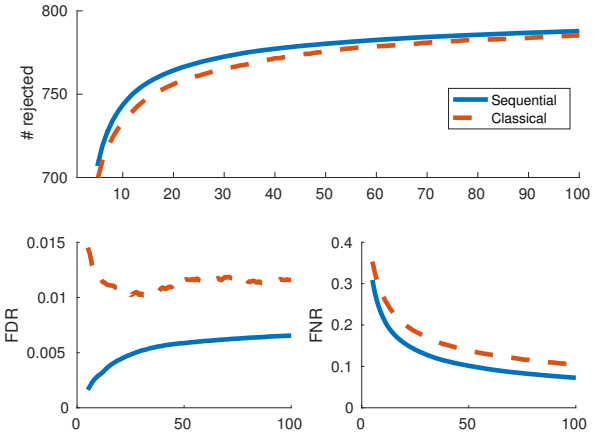

Figure 1: Experiments for sequential vs. fixed-horizon MHT, plotted against thousands of samples.

p-value in Corollary 4.2. The simulation results are presented in Fig. 1. The top plot shows the number of rejected tests per thousand sample points; the perfect algorithm (with unlimited data) would reject all 800 true alternatives. We see that the sequential does lead to more rejections, as desired. The FDR and FNR are plotted against budget in the bottom left and right plots, respectively. The sequential algorithm has better performance in both metrics: it makes fewer false positives and false negatives.

8 Conclusion and Future Work

We introduced a unified framework for sequential multiple hypothesis testing that includes most of the known common error notions such as Family-Wised Error Rate (FWER) and False Discovery Rate (FDR). After familiarizing the reader with sequential p-values and providing several examples, we showed how the MHT procedure can be successively applied to the sequential p-values to create a sequential procedure that dynamically allocates samples, stops as soon as enough evidence is obtained, and retains the same false positive control. We then discussed type II error and provided some experimental validation.

There are many interesting directions. Type II errors are still not well-understood beyond the simple vs. simple hypothesis case; what classes of hypotheses are compatible with guarantees on type II error? We have sacrificed performance for generality by using sequential p-values as our starting point; we would like to investigate when we can exploit the structure of specific sequential p-values to obtain tighter type I error guarantees. Lastly, our (f, q)-guarantee encompasses most known type I error criteria, but there likely exist looser notions that are more amenable to providing error guarantees in the sequential setting.

References

[1] Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 42–51, Arlington, Virginia, United States, 2016. AUAI Press.

[2] J. Bartroff and T. Lai. Multistage tests of multiple hypotheses. *Communications in Statistics-Theory and Methods*, 39(8-9):1597–1607, 2010.

[3] J. Bartroff and J. Song. Sequential tests of multiple hypotheses controlling false discovery and non-discovery rates. *arXiv:1311.3350*, 2013.

[4] J. Bartroff and J. Song. Sequential tests of multiple hypotheses controlling type I and II family-wise error rates. *Journal of statistical planning and inference*, 153:100–114, 2014.

[5] C. Baum and V. Veeravalli. A sequential procedure for multi-hypothesis testing. *IEEE Transactions on Information Theory*, 40(6), 1994.

[6] H. Baumgartner and C. Homburg. Applications of structural equation modeling in marketing and consumer research: A review. *International journal of Research in Marketing*, 13(2):139–161, 1996.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[8] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[9] C. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

[10] S. De and M. Baron. Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *Journal of Statistical Planning and Inference*, 142(7):2059–2070, 2012.

[11] V. Dragalin, A. Tartakovsky, and V. Veeravalli. Multihypothesis sequential probability ratio tests. ii. Accurate asymptotic expansions for the expected sample size. *IEEE Transactions on Information Theory*, 46(4):1366–1383, 2000.

[12] V. Draglia, A. Tartakovsky, and V. Veeravalli. Multihypothesis sequential probability ratio tests. I. Asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461, 1999.

[13] W. Dupont and W. Plummer. Power and sample size calculations: A review and computer program. *Controlled clinical trials*, 11(2):116–128, 1990.

[14] G. Fellouris and A. Tartakovsky. Almost optimal sequential tests of discrete composite hypotheses. *arXiv preprint arXiv:1204.5291*, 2012.

[15] Peter Grünwald. Safe probability. *arXiv preprint arXiv:1604.01785*, 2016.

[16] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

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

[18] R. Johari, L. Pekelis, and D. Walsh. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv:1512.04922*, 2015.

[19] R. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American statistical association*, 90(431):928–934, 1995.

[20] T. Lai. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Communications in Statistics-Theory and Methods*, 13(19):2355–2368, 1984.

[21] E. Lehmann and J. Romano. Generalizations of the familywise error rate. *Annals of statistics*, 33(3):1138–1154, 2005.

[22] S. Nakagawa. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045, 2004.

[23] A. Novikov. Optimal sequential multiple hypothesis tests. *arXiv:0811.1297*, 2008.

[24] H Robbins and D Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, pages 415–436, 1974.

[25] H. Robbins and D. Siegmund. A class of stopping rules for testing parametric hypotheses. In *Herbert Robbins Selected Papers*, pages 293–297. Springer, 1985.

[26] S. Sarkar. False discovery and false nondiscovery rates in single-step multiple testing procedures. *The Annals of Statistics*, pages 394–415, 2006.

[27] S. Sarkar. Step-up procedures controlling generalized FWER and generalized FDR. *The Annals of Statistics*, pages 2405–2420, 2007.

[28] D. Seugmund. *Sequential Analysis: tests and confidence intervals*. Springer Science & Business Media, 1985.

[29] G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and $p-$values. *Statistical Science*, pages 84–101, 2011.

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

[31] S. Zehetmayer, P. Bauer, and M. Posch. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, 21(19):3771–3777, 2005.

[32] S. Zehetmayer, P. Bauer, and M. Posch. Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in medicine*, 27(21):4145–4160, 2008.