Always Valid Inference: Bringing Sequential Analysis to A/B Testing

Ramesh Johari*
Leo Pekelis†
David J. Walsh‡

Abstract

A/B tests are typically analyzed via p-values and confidence intervals; but these inferences are wholly unreliable if users make decisions while *continuously monitoring* their tests. We define *always valid* p-values that let users try to take advantage of data as fast as it becomes available, providing valid statistical inference whenever they make their decision. Always valid p-values can be interpreted as the natural p-values corresponding to a sequential hypothesis test. Through this connection we derive always valid p-values with good detection properties. Notably, we also extend our approach to address multiple hypothesis testing in the sequential setting. Our methodology has been implemented in a large scale commercial A/B testing platform, from which we present empirical results.

1 Introduction

Technology platforms (such as web applications) typically optimize their product offerings using randomized controlled trials (RCTs), or *A/B testing*. A/B tests deliver *inference*: they control for exogenous factors that could influence observed outcomes, quantifying the differences between the variations being tested. The rapid rise of A/B testing has led to the emergence of a number of widely used platforms for implementation and analysis of experiments [7, 22].

These tools use standard frequentist statistical measures: p-values and confidence intervals. We begin with a reminder of how to interpret the p-value in a classical hypothesis testing framework. A standard A/B test with two variations (*control* and *treatment*) has a *null hypothesis* that both groups share the same parameter (e.g., customer

*Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, ramesh.johari@stanford.edu. RJ serves as an advisor to Optimizely, Inc.

†Department of Statistics, Stanford University, Stanford, CA 94305, lpekelis@gmail.com. LP was employed by Optimizely, Inc., when this work was carried out. He is also currently employed by Optimizely.

‡Department of Statistics, Stanford University, Stanford, CA 94305, dwalsh@stanford.edu. DJW was employed by Optimizely, Inc., when this work was carried out. He is also currently a part-time employee at Optimizely.

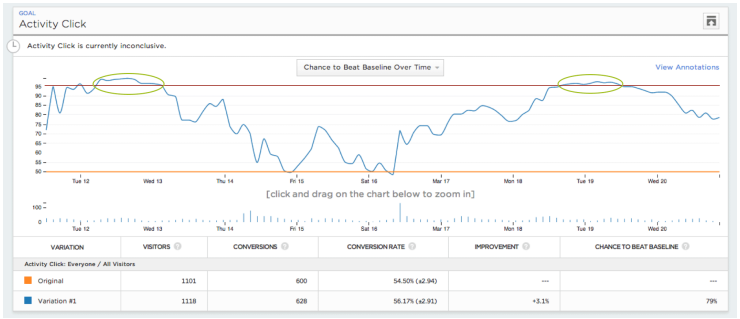

Figure 1: A typical dashboard from a large commercial A/B testing platform service. The graph depicts the "chance to beat baseline" of a test, which measures $1 - p_n$ over time, where p_n is the p-value after n observations of the null hypothesis that the clickthrough rate in treatment and control is identical. This particular test is a *A/A test*: both the treatment and control are the *same*. The graph shows that $1 - p_n$ rises above the 95% significance threshold if the user continuously monitors the test, triggering a Type I error.

conversion rate), and alternative that they are different. The p-value is then the probability of observing data as extreme as that observed, if the null hypothesis had been true.

It is worth noting that measures such as p-values have come under increasing scrutiny of late. For example, "p-value hacking" is a term given to the practice of data mining until statistically significant results are found, but not updating significance calculations to account for the search. P-values and confidence intervals can provide valid inference, but only when interpreted and used correctly.

Given this skepticism, why are p-values and confidence intervals so prevalent in these platforms? Their main benefit is that they provide *objective* measures of inference. P-values are powerful because they are *interpretable*: they give a common unit of statistical measurement of risk. For example, if a user rejects the null hypothesis when $p \leq \alpha$, they are guaranteed to have controlled their risk of false positives (i.e., Type I errors) at level α, without additional knowledge of the experiment itself. Equally important, p-values enable *transparency* across multiple observers of the same experiment: by reporting a valid p-value, interpretation of experimental output can be calibrated to the personal tolerance for error of each observer. The same arguments apply to confidence intervals.

However, these measures are objective *only if* properly used. Notably, they are computed under the assumption that the experimenter will not continuously monitor their test—in other words, there should be no "peeking" at the results [14]. Repeated significance testing (with rejection or continuation on the basis of those results) is a particularly pernicious form of "p-value hacking": it can lead to very high false positive probabilities—well in excess of the nominal α. In fact, as an extreme example, it can be shown that stopping the first time that $p_n \leq \alpha$ actually has Type I error of 100%

[20]. Even on moderate sample sizes (e.g., 10,000 samples, which is quite common in online A/B testing), Type I error can be inflated by over two-fold; see Appendix B. That is a problem both for the original user and for anyone else using the same p-value, since interpretability and transparency are lost.

Ultimately, users continuously monitor because it aligns with their incentives. They want to find true effects as quickly as possible, and technology has brought down the cost (see Figure 1 for an example of an A/B testing dashboard). In this paper, we claim *the user is right*: they should be able to use data as it arrives, and stop tests in a data-dependent manner. Thus we address the following challenge: *can we present users with the exact same simple dashboard, enabling continuous monitoring of p-values and confidence intervals, and yet guarantee valid inference?*

Dynamic monitoring of tests, and data-dependent rejection, place us squarely in the realm of *sequential hypothesis testing* [23, 20, 9, 19]. With that viewpoint, our main contributions are as follows. In Section 3, we implement p-values for sequential hypothesis tests; the analogous theory for confidence intervals is developed in Appendix A. Our definition is *always valid*: users can stop the test at any data-dependent time, and rejecting the null if the p-value is below α at that time controls Type I error (analogously for confidence intervals). We show that our definition is essentially tight.

Next, in Section 4, we discuss a particular sequential test, the *mixture sequential probability ratio test* (mSPRT) [17, 18, 16], and compare its performance to non-sequential testing. These results are novel as normal data with a normal prior are not covered by existing optimality literature (c.f. [8]). We also solve for optimal choice of mixing parameters, and present simulation results. *Together with the "user interface" of always valid p-values and confidence intervals, our solution provides the interpretability and transparency of standard statistical measures, with continuous monitoring, and faster results.*

The work in this paper has been deployed in a production A/B testing platform, serving thousands of clients worldwide. In Section 5 we apply our analysis to provide p-values and confidence intervals for A/B testing platforms. We also discuss extensions to procedures that adaptively change allocation over time (such as multiarmed bandits). In Section 6, we conclude with a discussion of error control for multiple sequential hypothesis tests; in particular, we demonstrate how *false discovery rate* (FDR) can be controlled in the sequential setting.

2 Preliminaries

To begin, we suppose that our data can be modelled as independent observations from an exponential family $\mathbf{X} = (X_i)_{i=1}^{\infty} \overset{iid}{\sim} F_\theta$, where the parameter θ takes values in $\Theta \subset \mathbb{R}^p$. Throughout the paper, $(\mathcal{F}_n)_{n=1}^{\infty}$ will denote the filtration generated by $(X_i)_{i=1}^{\infty}$ and $\mathbb{P}_\theta$ will denote the measure (on any space) induced under the parameter θ. Our focus is on testing a simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \neq \theta_0$. (In Section 5 we adapt our analysis to two-sample hypothesis testing, as is needed to test differences between control and treatment in an A/B test.)

Decision rules and sequential tests. In general, a decision rule is a mapping (T, δ) from sample paths $\mathbf{X}$ to a (possibly infinite) stopping time T that denotes the sample

size at which the test is ended, and a binary-valued, $(\mathcal{F}_T)$-measurable random variable δ that denotes whether or not H_0 was rejected. Decision rules where the terminal sample size may be data-dependent are commonly referred to as *sequential tests*.

Type I error. Type I error is the probability of erroneous rejection under the null, i.e., $\mathbb{P}_{\theta_0}(\delta = 1)$. Assuming that the user wants to bound Type I error, we will typically consider a family of decision rules parameterized by their Type I error rate $0 < \alpha < 1$. We assume these tests are *nested* in the following sense: $T(\alpha)$ is a.s. nonincreasing in α, and $\delta(\alpha)$ is a.s. nondecreasing in α. In other words, less stringent Type I error control allows the test to stop sooner, and is more likely to lead to rejection.

Fixed horizon testing. Under the default *fixed horizon* testing approach, we restrict to decision rules (n, δ), where the stopping time is required to be deterministic. In this setting, the objective is to maximize the power (the probability of detection under H_1) at that n. Indeed, for data in an exponential family, for any given n, there exist a family of uniformly most powerful (UMP) tests parameterized by α, which maximizes power uniformly over θ among tests with Type I error rate α. These tests reject the null if a particular test statistic τ_n exceeds a threshold $k(\alpha)$.

While this test maximizes power for the given n, the power increases as n is increased. The user must choose n to trade off power against the opportunity cost of waiting for more samples. Popular sample size calculators help. They ask for a "minimum detectable effect" (MDE), as well as a desired Type II error constraint β; the MDE is the smallest θ that the user would like to detect, with probability at least $1 - \beta$. Many standard statistics textbooks cover this procedure; see, e.g., [6].

The fixed horizon user interaction model. Testing platforms allow users to implement their optimal test via *p-values*. Specifically, the p-value at time n corresponding to the UMP test is:

$$p_n = \inf\{\alpha : \tau_n \geq k(\alpha)\}.$$

In other words, this p-value is the *smallest α such that the α-level test with sample size n rejects H_0*.

The interpretation is straightforward: p_n represents the chance of seeing a test statistic as extreme as τ_n under the null. Further, the process p_n provides sufficient information for the user to implement her desired test with ease: she waits for her chosen n, and rejects the null hypothesis if $p_n \leq \alpha$. In addition, p_n ensures transparency in the following sense: since each rule $\delta_n(\alpha)$ controls Type I error at level α, any other user can threshold the p-value obtained at her own appropriate α level to satisfy her desired Type I error bound.

In fact to control Type I error, we require only that the p-value is *super-uniform*:

$$\mathbb{P}_0(p_n \leq s) \leq s \text{ for all } s \in [0, 1]. \tag{1}$$

More generally, we refer to any $[0, 1]$-valued, $(\mathcal{F}_n)$-measurable random variable $\tilde{p}_n$ that satisfies (1) as a *fixed horizon p-value* for the choice of sample size n; we refer to the entire sequence $(\tilde{p}_n)_{n=1}^{\infty}$ as a *fixed horizon p-value process*. Of course p-values other than those defined above are rarely used in practice, because the associated decision rule $(n, \mathbf{1}\{\tilde{p}_n \leq \alpha\})$ has suboptimal power.

Given one rule $\delta_n(\alpha)$ for testing each θ_0, an $(1 - \alpha)$-level fixed-horizon confidence interval is the set of null values that are not rejected. With probability $(1 - \alpha)$, this

interval will capture the true parameter.

3 Always Valid Inference

Our goal is to let the user stop the test whenever they want, in order to trade off power with run-time as they see fit; the p-value they obtain should control Type I error. Our first contribution is the definition of *always valid* p-values as those processes that acheive this control:

Definition 1 (Always valid p-values)**.** *A fixed horizon p-value process* (p_n) *is* always valid *if given any (possibly infinite) stopping time* T *with respect to* $(\mathcal{F}_n)$*, there holds:*

$$\mathbb{P}_{\theta_0}(p_T \leq s) \leq s \quad \forall s \in [0,1]. \tag{2}$$

(An analogous definition is used to describe always valid confidence intervals; see Appendix A for details.)

The following theorem connects always valid p-values with the existing sequential testing literature. We leverage this connection in Section 4 to construct always valid p-values that can detect true effects effectively.

Theorem 1. *Let* $(T(\alpha), \delta(\alpha))_{\alpha \in [0,1]}$ *be a family of sequential tests. Then*

$$p_n = \inf\{\alpha : T(\alpha) \leq n, \delta(\alpha) = 1\}$$

defines an always valid p-value process.

Further, for any always valid p-value process $(p_n)_{n=1}^{\infty}$*, a sequential test* $(T(\alpha), \delta(\alpha))$ *is obtained from* $(p_n)_{n=1}^{\infty}$ *as follows:*

$$T(\alpha) = \inf\{n : p_n \leq \alpha\}; \tag{3}$$

$$\delta(\alpha) = \mathbf{1}\{T(\alpha) < \infty\}. \tag{4}$$

Note that (3)-(4) represent the most natural way for the user to make decisions based on always valid p-values: stop the first time that the p-value process hits α.

Proof. For the first result, nestedness implies the following identities for any $s \in [0,1], \varepsilon > 0$:

$$\{p_n \leq s\} \subset \{T(s+\varepsilon) \leq n, \delta(s+\varepsilon) = 1\} \subset \{\delta(s+\varepsilon) = 1\}.$$

$$\therefore \mathbb{P}_{\theta_0}(p_n \leq s) \leq \mathbb{P}_{\theta_0}(\delta(s+\varepsilon) = 1) \leq s + \varepsilon$$

and the result follows on letting $\varepsilon \to 0$. For the converse, we observe that for any $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}_{\theta_0}(\delta(\alpha) = 1) = \mathbb{P}_{\theta_0}(T(\alpha) < \infty) &\leq \mathbb{P}_{\theta_0}(p_{T(\alpha)} \leq \alpha + \varepsilon) \\ &\leq \alpha + \varepsilon\end{aligned}$$

where the last inequality follows from the definition of always validity. Again the result follows on letting $\varepsilon \to 0$. □

The p-value defined in Theorem 1 is not the unique always valid p-value associated with that sequential test. However, it is a.s. minimal among such always valid p-values for every n, resulting from the fact that it is uniquely a.s. *monotonically nonincreasing* in n. With a one-to-one correspondence between monotone always valid p-value processes and families of sequential tests, these processes can be seen as the natural representation of sequential tests in a streaming p-value format.

4 Optimal Sequential Tests

With duality between always valid p-values and sequential tests established, we answer the question: what is an optimal sequential test to use?

Since experimentation online is classified by plentiful data, and a dynamic environment, the user typically wants to run experiments as long as she wants, update preferences along the way, and obtain results quickly. We show the mixture Sequential Probability Ratio Test (mSPRT) achieves this through novel results comparing its expected run time to fixed horizon tests, and further optimize the mixture for fast detection.

The mSPRT is a well studied class of sequential tests first introduced in [16]. Consider the set of mixture likelihood ratios,

$$\Lambda_t^H(x) = \int_\Theta \frac{f_\theta(x)}{f_0(x)} dH(\theta) \tag{5}$$

where H is a mixing distribution over the parameter space Θ. The mSPRT is the sequential test with $T(\alpha) = \inf\{n : \Lambda_n^H(S_n) \geq \alpha^{-1}\}$ and $\delta_\alpha = 1(T(\alpha) < \infty)$, where $S_n = \sum_{i=1}^n X_i$. [1]

A key feature of the mSPRT is it is a *test of power one*, or $\mathbb{P}_\theta(T(\alpha) < \infty) = 1$ for all $0 < \alpha < 1$ and $\theta \neq 0$ [18]. This allows the user to wait essentially indefinitely to detect small effect sizes, or concede inconclusiveness through feedback from always valid p-values.

4.1 Fast Detection

We now show the mSPRT achieves *fast detection* over a prior. Most of the technical work to prove these results is contained in Theorem 6 in the Appendix. Compared to previous results such as [8], ours are novel in that they examine the truncated mSPRT, $T(\alpha) \wedge n$. This allows for analysis of a wider class of distributions, for example normally distributed observations with a normal prior, and direct comparison to traditional, fixed horizon testing.

For ease of exposition, we specialize to normal data, $X_i \sim N(\theta, 1)$, and $\theta_0 = 0$.[2] We also derive results as $\alpha \to 0$, and consequently, to have non-zero chance to reject,

[1] The choice of threshold α^{-1} on the likelihood ratio ensures Type I error is controlled at level α, via standard martingale techniques [20].

[2] It is possible to extend results to tests on the natural parameter of exponential families, as well as general priors that are positive and continuous on Θ. In fact, some misspecification of f_θ in (5) is permitted provided the mean as a function of θ is correct (see [12])

$n \to \infty$ in the fixed horizon. This is technically for tractability of the mSPRT, and practically, identifies the A/B tester seeking nearly certain results and having lots of data.

Recall the fixed horizon procedure is to determine a MDE level of θ, α, and β, to calculate the required $n(\theta, \alpha, \beta)$. A problem occurs when there is not enough information to get a good estimate of the MDE before starting the test: too low and the user is locked into lengthy experiments; too high and the effective power $1 - \beta$ plummets. We model this uncertainty as a normal prior over the effect size, $\theta \sim N(0, \tau)$.

Compare this to truncating a mSPRT at maximum size n_S, and admitting an inconclusive result if we ever reach it. This is a sequential test with $T'(\alpha) = T(\alpha) \wedge n_S$ and $\delta'(\alpha) = 0$ on the event $T(\alpha) = n_S$. The following proposition establishes the asymptotic Type II error probability, β, for both the truncated mSPRT and UMP fixed horizon test when $n \to \infty$ fast enough.

Proposition 1. *If $\alpha \to 0$, $n \to \infty$ such that $\log \alpha^{-1}/n \to 0$,*

$$\beta_k = \mathbb{E}_{\theta \sim N(0,\tau)} 1 - \beta_k(\theta) \sim C_k(\alpha) \frac{2\sqrt{2}}{\tau} \left(\frac{\log \alpha^{-1}}{n} \right)^{1/2}$$

where $k \in \{(f)ixed,\ m(S)PRT\}$, $0 < C_k(\alpha) < 1$ and

$$C_f(\alpha) = \int_0^1 \bar{\Phi}\left(\sqrt{\log \alpha^{-1}}(x - 1) \right) dx$$
$$C_S(\alpha) = \int_0^1 \bar{\Phi}\left(\sqrt{\frac{1}{2} \log \alpha^{-1}}(x^2 - 1) \right) dx,$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$, the right tailed standard Normal CDF.

It follows that similar asymptotic power can be achieved by $n_S = \left(C_S(\alpha)/C_f(\alpha)\right)^2 n$.

The UMP fixed horizon test has no choice but to wait for all of its samples, and so its sample size will always be n. Conversely, this is not true for the mSPRT; as the following Theorem shows, there is a clear benefit of stopping early.

Theorem 2. *Let T be the random time when the mSPRT for data $x_i \sim N(\theta, 1)$, and hypothesis $H_0 : \theta = 0$, first rejects at level α. Then for $\alpha \to 0$, $n \to \infty$ such that $\log \alpha^{-1}/n \to 0$,*

$$\mathbb{E}_{\theta \sim N(0,\tau)} \mathbb{E}_\theta (T \wedge n) = o(n)$$

up to terms which are $o(1)$.

(Explicit coefficients for the preceding result are given in the proof in Appendix C.)

Thus we find that the mSPRT can achieve similar asymptotic power to the fixed UMP test with average sample size,

$$\mathbb{E}_{\theta \sim N(0,\tau)} \mathbb{E}_\theta (T \wedge n_S) = o\left(n_S\right) = o(n), \tag{6}$$

of smaller order. This surprising superiority is precisely due to the ability of the mSPRT to automatically calibrate its sample size to the effect size of the test.

Yet these gains still understate the benefit from sequential testing. We have not incorporated the additional flexibility offered by the mSPRT—allowing the user to adjust her α, β, $\delta'(\alpha)$, and truncating n after the test has started; we intend to pursue analysis of this impact in future work.

4.2 Optimal choice of mixture

The following theorem is an immediate consequence of Theorem 6, along with equation (67) of [12].

Theorem 3. *Let X_i, $i = 1, \ldots,$ be drawn from an exponential family with density $f_\theta(x)$. Suppose H_γ is a parametric family, $\gamma \in \Gamma$, with density h_γ positive and continuous on Θ. Then up to $o(1)$ terms as $\alpha \to 0$, $\mathbb{E}_G E_\theta(T \wedge n)$ is minimized by*

$$\gamma^* \in \arg\min_{\gamma \in \Gamma} -\mathbb{E}_{\theta \sim G} \mathbf{1}_A I(\theta)^{-1} \log h_\gamma(\theta) \tag{7}$$

for $A = \{\theta \,:\, I(\theta) \geq (\log \alpha^{-1})/n\}$, $I(\theta) = \theta\Psi'(\theta) - \Psi(\theta)$, and Ψ the log-partition function for f_θ. Optimizing for h_γ does not impact first order terms of $\mathbb{E}_G E_\theta(T \wedge n)$.

Returning to our example of normal data, prior and now using a normal mixture, $h_\gamma(\theta) = \frac{1}{\gamma}\phi(\frac{\theta}{\gamma})$, the optimal choice of mixing variance becomes:

$$\tau^{2*} = \sigma^2 \frac{\Phi(-b)}{\frac{1}{b}\phi(b) - \Phi(-b)}. \tag{8}$$

This is a remarkably simple expression: it implies a rough *matching* of the mixing variance to the prior variance, with a correction for truncating. The correction tends to $\{0, \infty\}$ with b, showing that sampling efficiency is gained from focus on large effects when few samples are available, and smaller effects where there is ample data.

Simulation results support the theory: using (8) matches the average runtime minimizing mixture variance to within a factor of 10. There is also considerable robustness in choosing τ. A factor of 10 misspecification increases average runtime by less than 5%, while a factor of 1000 misspecification increases average runtime by a factor of 2. Table 1 in Appendix C provides more details.

Finally, we note that mixture-prior matching is important in practice as well. Our own numerical analysis of over $40,000$ historical A/B tests on a leading industry platform (Section 5) showed that prior matching has a significant beneficial effect on the run length of experiments.

5 Application to A/B Testing

The optimal sequential test described in the preceding section led to a commercial implementation (launched in January 2015) in a large scale platform serving thousands of clients, ranging from small businesses to large enterprises. In this section we describe how our results transfer to industrial practice, and we address some key challenges.

Two data streams. The primary challenge in applying our results to A/B testing is that there are two data streams: the control and the treatment. A common use case is that each of these streams consists of binary data (e.g., clicks, conversions, etc.). Formally, $\mathbf{X} = (X_i)_{i=1}^{\infty}$ where the X_i are i.i.d. Bernoulli(p_0); and $\mathbf{Y} = (Y_i)_{i=1}^{\infty}$ where the Y_i are i.i.d. Bernoulli(p_1), and we test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, where $\theta = p_1 - p_0$. It is useful to parameterize this composite null in terms of $\bar{p} = (p_0 + p_1)/2$.

We suppose that the data arrives in some possibly random order. If at a certain time, there are m and n observations from $\mathbf{X}$ and $\mathbf{Y}$ respectively, as before we base our choice of whether to stop and the terminal decision on the likelihood ratio for $\{X_1, ..., X_m, Y_1, ..., Y_n\}$. Throughout define $\mathbf{X}_{1:m} = (X_1, \ldots, X_m)$, $\mathbf{Y}_{1:n} = (Y_1, \ldots, Y_n)$, and

$$\mathsf{LR}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}); \bar{p}, \theta) = \frac{\mathbb{P}_{\bar{p},\theta}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}))}{\mathbb{P}_{\bar{p},0}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}))}.$$

Randomized controlled trials. In an RCT, allocation is made equally to the two groups. For simplicity suppose observations arrive in pairs $(m = n)$, so we can recover a hypothesis testing problem in terms of a single stream $Z_i = (X_i, Y_i)$. We address the composite null by treating $\bar{p}$ initially as known. Then we can use the mSPRT: we reject H_0 when

$$\Lambda_n^H = \int \mathsf{LR}(\mathbf{Z}_{1:m}; \bar{p}, \theta)\, dH(\theta) \tag{9}$$

exceeds $1/\alpha$. This test bounds Type I error at level α and has low average run-time.

There are two issues: (1) $\bar{p}$ is in fact unknown; and (2) even if it were known, this mSPRT is computationally challenging to implement in a streaming environment. By making a Central Limit Theorem approximation (see Lemma 2 in the Appendix), we solve both issues simultaneously. We approximate Λ_n^H with

$$\tilde{\Lambda}_n^H = \int \frac{\phi_{(\theta,\hat{V}_n)}(\bar{Y}_n - \bar{X}_n)}{\phi_{(0,\hat{V}_n)}(\bar{Y}_n - \bar{X}_n)}\, dH(\theta) \tag{10}$$

where $\phi(\mu, \sigma^2)$ is the density of a $N(\mu, \sigma^2)$ random variable.

This is easy to compute, at least when H is Gaussian. This approximation is good when n is large, so the hitting times of $1/\alpha$ for Λ_n^H and $\tilde{\Lambda}_n^H$ are similar when α is small. In particular, as $\tilde{\Lambda}_n^H$ does not depend on $\bar{p}$, we may test the composite hypothesis by thresholding it at $1/\alpha$. Since this test approximates the exact mSPRT for any $\bar{p}$, it provides Type I error control uniformly over the composite null, and further it has low average run-time under H_1 for any true $\bar{p}$ (provided the prior for θ is calibrated appropriately).

Adaptive allocation and bandits. More generally, allocation to treatment and control can be *adaptive*. For example, a multiarmed bandit policy may allocate observations to treatment and control to optimally tradeoff exploration of the two variations against exploitation of the better performing variation [10, 1]. When $\bar{p}$ is known there is a natural extension of the mSPRT used above to the bandit setting. The following theorem shows that indeed it controls Type I error. In particular, this result allows us to generate *always valid p-values for adaptive allocation strategies* in that case.

Theorem 4. *Consider an arbitrary allocation rule, where the processes $m(t), n(t)$ represent the number of observations on each stream among the first t total observations. Let*

$$\Lambda_t^H = \int \mathsf{LR}(\mathbf{X}_{1:m(t)}, \mathbf{Y}_{1:n(t)}; \bar{p}, \theta)\, dH(\theta) \tag{11}$$

Then the test which rejects H_0 as soon as $\Lambda_t^H \geq 1/\alpha$, controls Type I error at level α.

The proof rests on the fact that Λ_t^H is a martingale. This is well-known for deterministic allocation; the property extends to bandits because the arm pulled at $t+1$ is conditionally independent of the value observed, given the first t observations. The desired bound on the hitting probabilities then follows by the Optional Stopping Theorem.

Unfortunately, Λ_t^H may depend heavily on the value of $\bar{p}$, so it is more challenging to extend to the case of $\bar{p}$ unknown. A natural approach is to estimate this value by $\hat{\bar{p}} = (\bar{X}_{m(n)} + \bar{Y}_n)/2$ and then approximate Λ_t^H by

$$\hat{\Lambda}_t^H = \int \mathsf{LR}(\mathbf{X}_{1:m(t)}, \mathbf{Y}_{1:n(t)}; \hat{\bar{p}}, \theta)\, dH(\theta) \tag{12}$$

but we do not how much this may inflate the Type I error. This issue, as well as extending the results to general multiarmed bandits and best arm selection problems, remain important directions for future work.

Empirical results. In Figure 2, we carried out the following numerical experiment using data from an industry leading A/B testing platform. For over 40,000 client experiments, we first estimated the effect size θ using the observed effect size in the experiment, and then computed an "optimal" fixed horizon sample size for a minimum detectable effect (MDE) of θ. In other words, we compute the sample size needed for Type I error control of 90% and Type II error control of 20%, i.e., power of 80%, with an MDE of θ. We compared this sample size to the run length of our mSPRT, with the same Type I error control of 90%.

Figure 2 shows the ratio of these run lengths. As can be seen from the dotted line, there is a penalty paid for the flexibility of a sequential test. The sequential test sometimes runs shorter than the fixed horizon test, but can also be 2x-3x longer in sample size.

However, now suppose that we misestimated the true effect size when computing our fixed horizon sample size. Indeed, we see that the sequential test is significantly more robust (dashed and solid lines): when the MDE is off by 30% or 50%, the sequential test run lengths are typically much lower than the fixed horizon sample size. Note that an MDE off by 30%-50% in practice is quite likely, given that differences in conversion or clickthrough rates are rarely more than a few percentage points. In this sense, our sequential testing approach yields significant sample size benefits to the user, despite more disciplined control of Type I error.

6 Multiple Testing

When multiple experiments are conducted simultaneously, always valid p-values and confidence intervals can be derived for each test individually. However, their per-

experiment Type I error control may be insufficient, as the impact of multiple Type I errors over successive experiments may compound dramatically. *Multiple hypothesis testing* aims to bound a global error constraint: a function of the per-experiment Type I errors that better represents their combined cost to the user. In this section we extend our results to this setting; the methods presented have been commercially deployed in the same A/B testing platform described in the preceding section.

We focus on the two error functions most studied in the literature. The first function is the *family-wise error rate* (FWER):

$$\mathsf{FWER} = \max_{\boldsymbol{\theta}} \mathbb{P}_{\boldsymbol{\theta}}(\delta_i = 1 \text{ for at least one } i \text{ s.t. } \theta^i = \theta_0^i).$$

This is the worst-case probability of incurring any false positive; bounding FWER may be appropriate when the performance of two variations is assessed on multiple metrics and improvement on all metrics is required. The second is the *false discovery rate* (FDR):

$$\mathsf{FDR} = \max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}} \left\{ \frac{\#\{1 \leq i \leq m : \theta^i = \theta_0^i, \delta_i = 1\}}{\#\{1 \leq i \leq m : \delta_i = 1\} \wedge 1} \right\}.$$

This may be appropriate when the user wants to review many experiments to plan future optimization: if her plan is weighted equally towards every result where a non-zero effect is detected, the cost is simply the proportion of such conclusions that are false.

In fixed-horizon statistics, established procedures exist to bound these quantities. They take as input the entire vector of fixed-horizon p-values and produce as output a set of rejections, such that an objective is controlled. These procedures satisfy a transparency property that extends the transparency of fixed-horizon p-values described

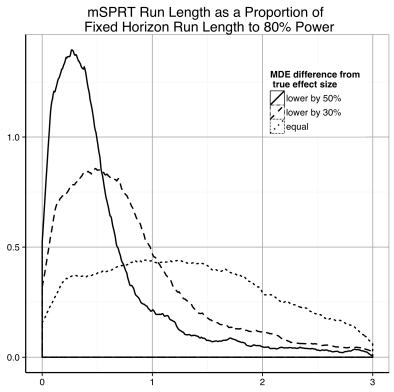

Figure 2: This figure shows the performance of our sequential test on data from over 40,000 experiments on a commercially deployed A/B testing platform. When the effect size is misestimated, our test provides better performance than the usual fixed horizon test. See Section 5 for detailed discussion.

in the Introduction. With multiple hypotheses, users may disagree on the global Type I error to be controlled, not just the level; nevertheless, different users can apply their chosen multiple testing correction to the same set of fixed-horizon p-values to achieve their desired Type I error control.

In the sequential context, we seek multiple testing procedures that map always valid p-values to a set of rejections, such that at any stopping time the desired error quantity is controlled. On sample paths where the stopping time is infinite, we treat the error incurred as zero. As ever, the purpose is to preserve transparency: no matter how the primary user chooses this stopping time, this enables a new user to apply their chosen multiple testing procedure to the p-values to obtain their desired Type I error control. If a multiple testing procedure satisfies this property for a class of always valid p-values, we say it *commutes with always validity* on that class.

The standard procedure to control the FWER in the fixed-horizon context is the Bonferroni correction [5]: this rejects hypotheses $(1), ...(j)$ where j is maximal such that $p^{(j)} \leq \alpha/m$, and $p^{(1)}, ..., p^{(m)}$ are the p-values arranged in increasing order. For FDR, the standard procedure is Benjamini-Hochberg [2]. Two versions of BH are used, depending on whether the data are known to be independent across experiments. If independence holds (BH-I), we reject hypotheses $(1), ..., (j)$ where j is maximal such that $p^{(j)} \leq \alpha j/m$; in the general (BH-G), we choose the maximal j such that:

$$p^{(j)} \leq \frac{\alpha j}{m \sum_{r=1}^{m} 1/r}.$$

The next two propositions show that Bonferroni and the general form of BH commute with always validity.

Proposition 2. *Let $(p_n^i)_{i=1}^m$ be always valid p-values, and let T be an arbitrary stopping time. Then the set of decisions obtained by applying Bonferroni to $\mathbf{p}_T$ controls FWER at level α.*

Proposition 3. *Let $(p_n^i)_{i=1}^m$ be always valid p-values, and let T be an arbitrary stopping time. The set of decisions obtained by applying the BH-G procedure to $\mathbf{p}_T$ controls FDR at level α.*

BH-I does not commute with always validity over all independent p-value processes. For a counter-example, let p_n^1 be a.s. constant across n with $p_1^1 \sim U(0,1)$, and let $p_1^2 = 1, p_n^2 = 0$ for $n \geq 2$. These are feasible always valid p-value processes when the 1st hypothesis is null and the 2nd is non-null. Consider the following stopping time: $T = 1$ if $p_1^1 \leq \alpha/2$, else $T = 2$. BH under independence applied to (p_T^1, p_T^2) gives an FDR of $3\alpha/2$. We note though that this example is somewhat artificial, because it leverages knowledge of a rejected null hypothesis as unfavorably as possible.

FDR control with BI-I does hold if some assumptions are placed on the stopping time. To analyse this, we begin by defining some stopping times associated with the p-values that play a key role in our analysis.

Definition 2. *Given independent always valid p-values $\mathbf{p}_n$, let S_n^{BH} be the rejections*

when BH-I is applied to these at level α and let $R_n^{BH} = |S_n^{BH}|$. Define:

$$T_r = \inf\{t \ : \ R_t^{BH} = r\};$$
$$T_r^+ = \inf\{t \ : \ R_t^{BH} > r\};$$
$$T_r^i = \inf\{t \ : \ p_t^i \leq \frac{\alpha r}{m}\}.$$

Now, if $p_{(1),n}^{-i}, p_{(2),n}^{-i}, ...$ are the p-values for the experiments other than i placed in ascending order, consider a modified BH procedure that rejects hypotheses $(1), ..., (k)$ where k is maximal such that $p_{(k),n}^{-i} \leq \alpha(k+1)/m$. Define the rejection set $(S_n^{BH})_0^{-i}$ as those obtained under the original BH-I procedure if $p_n^i = 0$. Let $(R_n^{BH})_0^{-i} = |(S_n^{BH})_0^{-i}|$ and define:

$$(T_r)_0^{-i} = \inf\{t \ : \ (R_n^{BH})_0^{-i} = r\}$$
$$(T_r^+)_0^{-i} = \inf\{t \ : \ (R_n^{BH})_0^{-i} > r\}.$$

We have the following theorem.

Theorem 5. *Given a stopping time T, let m_0 be the number of truly null hypotheses and let I be the set of null hypotheses i such that:*

$$\sum_{r=1}^{m} \mathbb{P}\left((T_{r-1})_0^{-i} \ \leq \ T < (T_{r-1}^+)_0^{-i} \ \middle| \ T_r^i \leq T \ , \ T < \infty\right) > 1 \tag{13}$$

Then the rejection set S_T^{BH} has FDR at most

$$\alpha\left(\frac{m_0}{m} + \frac{|I| \sum_{k=2}^{m} \frac{1}{k}}{m}\right).$$

In particular, if we permit only stopping times where I is empty, BH-I commutes with always validity and controls FDR.

The theorem provides a method to study FDR control for several natural stopping times that might be employed by the user. For example, perhaps the most natural stopping time for a user is the first time some fixed number $x \leq m$ hypotheses are rejected. Another very natural stopping time is the first time that the p-value on a fixed hypothesis crosses a threshold. For both stopping times, it can be shown that I is indeed empty. We prove these results in Appendix E. In addition, there we discuss stopping times where the user waits for rejection on a specific subset of hypotheses; this is a case where I may not be empty.

q-values. Although it is maximally transparent to let the user apply any multiple testing procedure, it requires reasonable statistical savviness on their part, as the set of p-values has no simple interpretation until the correction is applied. If there is no ambiguity in the multiple testing quantity to be controlled, with users differing only on their desired level, this can be addressed with a set of m *q-values* [21]. Just as users in the single hypothesis case obtain their desired Type I error control by thresholding

the p-value, users may obtain their desired global error control by thresholding each q-value. For details, see Appendix E.

Confidence intervals. Effective decision-making may also require confidence intervals that satisfy some global coverage bound. Analogous to controlling FWER, the user may wish to bound the probability that any confidence interval fails to contain the true value. Just like the Bonferroni correction for p-values, implementing always valid confidence intervals at level $(1 - \alpha/m)$ bounds FWER uniformly over all stopping times.

Similar to FDR, we often want confidence intervals that control the *False Coverage Rate (FCR)* [4]: the expected proportion of intervals that fail to contain the true parameter among a subset of experiments that have been selected by some data-dependent rule. FCR control is nontrivial in our setting, because the selection rule is an *unknown property of the user*. Whereas users typically only view the p-values of significant experiments, they may wish to gauge the range of plausible parameter values even on tests i where H_0^i is not rejected.

We approximate the selection rule as the union of the significant experiments and some fixed set J of experiments, with $j = |J| \ll m$, which are always of interest to the user. Then the next proposition follows the approach of Theorem 1 in [4] to achieve approximate FCR control in the fixed-horizon context. The idea is that it is the aggressive selection rules, which choose few experiments, that can obtain the highest FCR. Where that paper requires the selection rule to be known, we are conservative by reducing the nominal significance level of the CIs by an underestimate of the proportion of experiments that are selected.

Proposition 4. *Given fixed-horizon p-values* $\mathbf{p}$*, let* S^{BH} *be the rejection set under BH-I,* $R^{BH} = |S^{BH}|$*, and* $(CI^i(1-s))_{i=1}^m$ *be the corresponding fixed-horizon CIs at each level* $s \in (0, 1)$*. Define the corrected confidence intervals:*

$$\tilde{CI}^i = \begin{cases} CI^i(1 - R^{BH}\alpha/m) & i \in S^{BH}; \\ CI^i(1 - (R^{BH}+1)\alpha/m) & i \notin S^{BH}. \end{cases} \tag{14}$$

Then for any J*, if the selection rule is the experiments* $J \cup S_{BH}$*, the FCR is at most* $\alpha(1 + j/m)$*.*

In Appendix E, we show that this approximate FCR control is preserved in the sequential setting, if we make restrictions on the stopping time that are analogous to requiring $I = \varnothing$ for FDR control in applying Theorem 5.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[2] 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.

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

[4] Y. Benjamini and D. Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[5] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[6] A. Gelman and J. Hill. *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press Cambridge, 2007.

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

[8] T. L. Lai. On optimal stopping problems in sequential hypothesis testing. *Statistica Sinica*, 7(1):33–51, 1997.

[9] T. L. Lai. Sequential aanalysis: Some classical problems and new challenges. *Statistica Sinica*, 11:303–408, 2001.

[10] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[11] T. L. Lai and D. Siegmund. A nonlinear renewal theory with applications to sequential analysis i. *The Annals of Statistics*, 5(5):946–954, 09 1977. doi: 10.1214/aos/1176343950. URL `http://dx.doi.org/10.1214/aos/1176343950`.

[12] T. L. Lai and D. Siegmund. A nonlinear renewal theory with applications to sequential analysis ii. *The Annals of Statistics*, 7(1):60–76, 01 1979. doi: 10.1214/aos/1176344555. URL `http://dx.doi.org/10.1214/aos/1176344555`.

[13] T. L. Lai and J. Q. Wang. Asymptotic expansions for the distributions of stopped random walks and first passage times. *The Annals of Probability*, pages 1957–1992, 1994.

[14] E. Miller. How not to run an A/B test. 2010. URL http://www.evanmiller.org/how-not-to-run-an-ab-test.html. Blog post.

[15] M. Pollak and D. Siegmund. Approximations to the expected sample size of certain sequential tests. *The Annals of Statistics*, 3(6):1267–1282, 11 1975. doi: 10.1214/aos/1176343284. URL http://dx.doi.org/10.1214/aos/1176343284.

[16] H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, pages 1397–1409, 1970.

[17] H. Robbins and D. Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429, 10 1970. doi: 10.1214/aoms/1177696787. URL http://dx.doi.org/10.1214/aoms/1177696787.

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

[19] D. Siegmund. Estimation following sequential tests. *Biometrika*, 65(2):341–349, 1978.

[20] D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 1985.

[21] J. D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035, 2003.

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

[23] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Supplementary Material

A Confidence Intervals

All the theory developed in this paper for always valid p-values carries over to confidence intervals as well. In this section we develop the definition of always valid confidence intervals. We note that construction of confidence intervals on completion of a sequential test has been studied [19]. However, the real-time interface we present for streaming confidence intervals is a crucial innovation. We begin with the following definition of a fixed horizon confidence interval, analogous to fixed horizon p-values.

All the theory developed in this paper for always valid p-values carries over to confidence intervals as well. In this section we develop the definition of always valid confidence intervals. We note that construction of confidence intervals on completion has been studied [19]. However, the real-time interface we present for streaming confidence intervals is a crucial innovation. We begin with the following definition of a fixed horizon confidence interval, analogous to fixed horizon p-values.

Definition 3 (Fixed horizon confidence interval)**.** *A* (fixed-horizon) $(1-\alpha)$-level confidence interval *(CI) process for θ is a process $(\mathsf{CI}_n)_{n=1}^{\infty} \in \mathcal{P}(\Theta)$ that is measurable wrt $(\mathcal{F}_n)$, such that for each n:*

$$\mathbb{P}_{\theta_0}(\theta_0 \in \mathsf{CI}_n) \geq 1-\alpha. \tag{15}$$

For such a process we define:

$$\mathsf{CI}_\infty = \liminf_{n\to\infty} \mathsf{CI}_n. \tag{16}$$

By analogy to always valid p-values, we can define always valid confidence intervals by extending the definition above to hold at any stopping time T.

Definition 4 (Always valid confidence interval)**.** *A fixed-horizon $1-\alpha$-level confidence interval process (CI_n) is* always valid *if, given any (possibly infinite) stopping time T with respect to $(\mathcal{F}_n)$, there holds:*

$$\mathsf{P}_{\theta_0}(\theta_0 \in \mathsf{CI}_T) \geq 1-\alpha. \tag{17}$$

We recall that in a fixed-horizon framework, p-values and CIs are "inverses" of each other in an appropriate sense: θ_0 lies in a $1-\alpha$ CI if and only if the p-value for the test of $\theta = \theta_0$ is greater than or equal to α. (The test of $\theta = \theta_0$ is the test of the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$.) The same duality holds for always valid p-values and CIs, as we show in the following proposition.

Proposition 5. *1. Suppose that, for each $\theta^* \in \Theta$, $(p_n^{\theta^*})$ is an always valid p-value process for the test of $\theta = \theta_0$. Then*

$$\mathsf{CI}_n = \left\{\theta^* : p_n^{\theta^*} > \alpha\right\}$$

is an always valid $(1-\alpha)$-level CI for θ.

2. *Conversely suppose that, for each* $\alpha \in [0,1]$, (CI_n^α) *is an always valid* $(1-\alpha)$*-level CI for* θ*. Then*

$$p_n = \inf\{\alpha : \theta_0 \in \mathsf{CI}_n^\alpha\}$$

is an always valid p-value process for the test of $\theta = \theta_0$*.*

Proof. Let T be a stopping time. We have:

$$\mathbb{P}_{\theta_0}(\theta_0 \in \mathsf{CI}_T) = \mathbb{P}_{\theta_0}(p_T^{\theta_0} > \alpha) \geq 1 - \alpha.$$

For $\alpha > t$, $\mathbb{P}_{\theta_0}(p_T < \alpha) = \mathbb{P}_{\theta_0}(\theta_0 \notin \mathsf{CI}_T^\alpha) < \alpha$. The result follows on letting $\alpha \to t$. □

As with always valid p-values, we can easily construct always valid confidence intervals using a nested family of sequential tests $(T(\alpha), \delta(\alpha))$ that each control Type I error at level α. Specifically, at time n, include $\theta_0 \in \mathsf{CI}_n$ if the sequential test for $\theta = \theta_0$ has $T(\alpha) \leq n$. These confidence intervals share the same interpretability and transparency benefits as always valid p-values: the user can choose to stop the test at a time that is optimal for their own decision problem, and the resulting confidence intervals yield $1 - \alpha$ coverage regardless of the stopping time chosen.

B Continuous Monitoring and Type I Error

As noted in the main text, continuous monitoring of fixed horizon p-values can lead to severely inflated Type I error. Theoretically, it is known that repeated significance testing — and in particular, stopping a test the first time the p-value drops below a fixed α — can lead to Type I error of 100%, assuming arbitrarily large sample sizes are allowed. In this section we investigate this result numerically, on finite sample sizes.

Specifically, we consider the following procedure. Suppose that the data is generated from a $N(0,1)$ distribution, and we test the null hypothesis $\mu = 0$ against the alternative that $\mu \neq 0$. We use a standard z-test, and let p_n denote the resulting p-value after n observations.

We consider a user who continuously monitors the test, and rejects the null if the p-value ever drops below α, for $\alpha = 0.1, 0.05, 0.01$. In addition, we consider a *post-hoc power* policy that works as follows: the user tracks the p-value, and rejects the null if the p-value drops below level $\alpha = 0.05$, *and* a sample size calculator (with power $1 - \beta = 0.8$) using the currently observed effect size at the MDE yields a required sample size which is smaller than the current number of observations. In other words, informally, the post hoc power calculation treats the currently observed effect as the true effect, and rejects if enough observations have been seen for detection at that effect size. This approach is commonly used in practice, but turns out to be equivalent to rejection at a fixed α.

In Figure 3 we show the Type I error for each of these policies, as a function of the sample size. It is clear that Type I error is severely inflated above the nominal Type I error control, higher than 2x even for the post-hoc power approach described above with 10,000 observations. For many web applications, 10,000 observations in an A/B test would be quite common. As a result, the plots demonstrate that the asymptotic Type

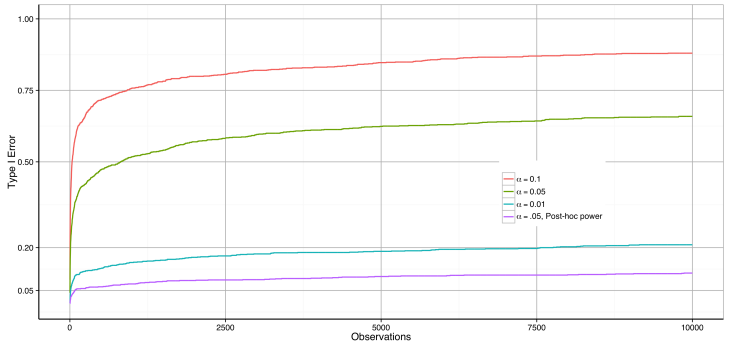

Figure 3: Type I error of repeated significance testing in finite sample sizes.

I error of 100% for repeated significance testing is a significant problem in practice as well, even on finite sample sizes.

C Proofs and Simulations for Optimal Sequential Test

To begin, we reiterate the setup in more detail.

Let $X_i \sim f_\theta, i = 1, 2, \ldots,$ be drawn i.i.d and real valued, with $f_\theta(x) = e^{\theta x - \Psi(\theta)} f_0(x)$, and $\theta \in \Theta$, an open interval on the real line. If f_0 is a measure on $\mathbb{R}$, standard properties of exponential families are that $\mathbb{E}[X] \equiv \mu = \Psi^{'}(\theta)$, and $\mathrm{Var}(X) \equiv \sigma^2 = \Psi^{''}(\theta)$. The relative entropy of θ_0 from θ is defined as $I(\theta, \theta_0) = (\theta - \theta_0)\Psi^{'}(\theta) - (\Psi(\theta) - \Psi(\theta_0))$. Since any exponential family may be re-centered to have $f_0 = f_{\theta_0}$, we present all our results without loss of generality for $\theta_0 = 0$, and $I(\theta) = I(\theta, 0)$. Finally we use the notation: $\mathbf{1}_A$ as the indicator of a set A, $\bar{A}$ as the complement of A, and $\log^2(x) = \log\log(x)$.

The set of mixture likelihood ratios is then

$$\Lambda_t^H(x) = \int_J \exp[(\theta - \theta_0)x - t(\Psi(\theta) - \Psi(\theta_0))] dH(\theta)$$

where H is a distribution with density h, assumed everywhere continuous and positive on Θ. The mSPRT for the null hypothesis $H_0 : \theta = \theta_0$ is a test $(T(\alpha), \delta(\alpha))$ with $T(\alpha) = \inf\{n : \Lambda_n^H(S_n) \geq \alpha^{-1}\}$, $\delta_\alpha = 1(T(\alpha) < \infty)$, and $S_n = \sum_{i=1}^n X_i$. We also assume the true parameter θ follows a prior distribution, $\theta \sim G(\theta)$, which is absolutely continuous with respect to the Lebesgue measure on Θ

The following theorem is the main technical contribution for this section. Motivating arguments, as well as background exposition may be found in [15].

Theorem 6.

$$\int E_\theta(T \wedge n) dG(\theta) = nPr_{G(\theta)}\left(\bar{A}\right) + E_{\theta \sim G}\left\{\mathbf{1}_A E_\theta(T)\right\} \tag{18}$$

up to $o(1)$ *terms as* $\alpha \to 0$.

Proof. Choose any $0 < \varepsilon < 1$, and $\delta > 0$. Define two times, $n_1 = (1-\varepsilon)log(\alpha^{-1})/I(\theta)$, $n_2 = (1+\varepsilon)log(\alpha^{-1})/I(\theta)$, $A_i = \{n \geq n_i\}$. On the set $\bar{A}$, we have

$$E_\theta(T \wedge n) = nP_\theta(T > n_1) + E_\theta(T \wedge n1_{T \leq n_1})$$

By lemma 1, for some $\lambda > 0$, $P(T \leq n_1) = O(\alpha^\lambda)$, and so

$$E_\theta(T \wedge n) = n + O(n\alpha^\lambda).$$

On the set A, we have

$$E_\theta(T \wedge n) = E_\theta(T) + \int_{n_2 > T \geq n} (n - T)dP_\theta - \int_{T > n_2} TdP_\theta.$$

By lemma 5 of [15], $T \sim \log(\alpha^{-1})/I(\theta)$, and $E_\theta T^2 \sim (\log(\alpha^{-1})/I(\theta))^2$. Note also that

$$\mathbb{E}_{\theta \sim G} \mathbf{1}_A I(\theta)^{-1} \leq n \Big/ \log \alpha^{-1} < \infty \tag{19}$$

And so using Cauchy-Schwartz, (19), and lemma 1,

$$\int_{T > n_2} TdP_\theta \leq \left(E_\theta(T^2)P_\theta(T \geq n_2)\right)^{1/2} = O(\alpha^{-\lambda} \log \alpha^{-1}).$$

Finally, let $B_\theta = \{\theta \ : \ I(\theta) \in [n_2/n, n_1/n]\}$. Then

$$\left| E_{G(\theta),A} \int_{n_2 > T \geq n} (n - T)dP_\theta \right| \leq P_{G(\theta)}(B_\theta) \sup_{\theta \in B_\theta} P_\theta(n_2 > T \geq n) \sup_{n_2 > t \geq n} |n - t| \leq P_{G(\theta)}(B_\theta) \frac{2n\varepsilon}{1-\varepsilon}.$$

The result follows since ε is arbitrary. □

Lemma 1. *For any* $0 < \varepsilon < 1$*, there exists a* $\lambda = \lambda(\theta) > 0$ *such that*

$$P_\theta(|T - \log(\alpha^{-1})/I(\theta)| > \varepsilon \log(\alpha^{-1})/I(\theta)) = O(\alpha^\lambda)$$

.

Proof. The case for $T < \log(\alpha^{-1})/I(\theta)$ is handled by Lemma 3 of [15]. For $T \geq \log(\alpha^{-1})/I(\theta)$, using Jensen's inequality

$$\begin{aligned} \log L(s_n, n) &= n \log \int e^{\theta s_n - n\psi(\theta)} G(d\theta) + \log \left\{ \frac{\int e^{\theta n(s_n/n - \psi(\theta))} G(d\theta)}{\left(\int e^{\theta s_n/n - \psi(\theta)} G(d\theta)\right)^n} \right\} \\ &\geq n \log \int e^{\theta s_n - n\psi(\theta)} G(d\theta) = n\gamma(s_n/n) \end{aligned}$$

and γ is a smooth, positive function. Hence $P_\theta(T(\alpha) > n) \leq P_\theta(T'(\alpha) > n)$ for

$$T'_n = \inf\{n \ : \ n\gamma(s_n/n) \geq \log(\alpha^{-1})\}.$$

It follows that for some $\delta > 0$, (c.f. lemma 6 of [13] and lemma 2 of [15])

$$\begin{aligned} P_\theta(T_\alpha > (1+\varepsilon)\log(\alpha^{-1})/I(\theta)) &\leq P_\theta(n = \frac{\log(\alpha^{-1})}{\gamma(\mu)+\varepsilon}\ ,\ |\gamma(s_n/n) - \gamma(\mu)| > \varepsilon) \\ &\leq P_\theta(n = \frac{\log(\alpha^{-1})}{\gamma(\mu)+\varepsilon}\ ,\ |s_n/n - \mu| > \max(\delta, \frac{\varepsilon}{2\|\nabla\gamma(\mu)\|})) \end{aligned}$$

The result follows since s_n/n has exponentially fast convergence to μ . □

Theorem 3 now follows immediately.

Proof of Theorem 3 . Clearly $0 \leq E_G E(T \wedge n) < \infty$ for all finite n. By Theorem 6, and equation (67) of [12], the terms involving H in $E_G E(T \wedge n)$ as $\alpha \to 0$ are

$$\begin{aligned} &- 2\mathbb{E}_{\theta\sim G}\mathbf{1}_A I(\theta)^{-1}\log h_\gamma(\theta) + o(1) \\ &= o\left(K\mathbb{E}_{\theta\sim G}\mathbf{1}_A I(\theta)^{-1}\log\alpha^{-1}\right) \end{aligned}$$

since $0 < h_\gamma(\theta) < \infty$ on Θ. □

We next specialize to the case of $f_\theta(x) = \phi(x)$ and prove the results in section 4.

Proof of Proposition 1. First, we prove the proposition for the fixed horizon test. Standard results show that $z_{1-2\alpha} \sim \sqrt{2\log\alpha^{-1}}$ as $\alpha \to 0$ so that

$$\begin{aligned} &1 - \beta_{fixed}(\theta) \\ &= \bar{\Phi}\left(|\theta|\sqrt{n} - z_{1-2\alpha}\right) = \bar{Phi}\left((\log\alpha^{-1})^{1/2}A_f(\theta,\alpha,n)\right), \end{aligned}$$

where $A_f(\theta,\alpha,n) = |\theta|\left(\frac{\log\alpha^{-1}}{n}\right)^{1/2} - \sqrt{2}$. Hence it is necessary to have $n/\log\alpha^{-1} \to \infty$ for $\beta_{fixed}(\theta) \to 1$. Define $B_f = \{\theta\ :\ A_f(\theta,\alpha,n) \geq \sqrt{2}\}$ and split the integral

$$\begin{aligned} \mathbb{E}_{\theta\sim N(0,\tau)}1 - \beta_{fixed}(\theta) &= \int_{B_f} 1 - \beta_{fixed}(\theta)\frac{1}{\tau}\phi(\theta/\tau)d\theta \\ &\quad + \int_{\bar{B}_f} 1 - \beta_{fixed}(\theta)\frac{1}{\tau}\phi(\theta/\tau)d\theta \\ &= (i) + (ii) \end{aligned}$$

where $\bar{B}_f$ is the complement of B_f and $\phi(x)$ the standard normal density. For $\theta \in B_f$, the standard tail bound on the Normal CDF, $\bar{\Phi}(x) \leq x^{-1}\phi(x)$ gives

$$\begin{aligned} \bar{\Phi}\left((\log\alpha^{-1})^{1/2}A_f(\theta,\alpha,n)\right) &\leq (4\pi\log\alpha^{-1})^{-1/2}\alpha \\ &= o(\alpha), \end{aligned}$$

so that $(i) = o(\alpha)$ as well. For term (ii), note that $\bar{B}_f \to \{0\}$ so that $\phi(\theta/\tau) \sim 1$. This, the change of variable $x = \left(\frac{2\log\alpha^{-1}}{n}\right)^{-1/2}\theta$ and symmetry of the integrand give

$$(ii) \sim \frac{2\sqrt{2}}{\tau}\left(\frac{\log\alpha^{-1}}{n}\right)^{1/2}\int_0^2 \bar{\Phi}\left((\log\alpha^{-1})^{1/2}(x-1)\right)dx.$$

The result follows on noting $\bar{\Phi}\left((\log\alpha^{-1})^{1/2}(x-1)\right) = o(1)$ when $x > 1$. For the mSPRT, we use the normal approximation to $P_\theta(T > n)$ from Theorem 2 of [11] which, in the case of standard normal data, gives

$$P_\theta(T > n) \sim \bar{\Phi}\left\{(\log\alpha^{-1})^{1/2}A_s(\theta,\alpha,n)\right\}$$

where $A_f(\theta,\alpha,n) = \frac{1}{2\sqrt{2}}\left(\theta^2\frac{\log\alpha^{-1}}{n} - 2\right)$. The rest of the proof proceeds as for the fixed horizon test, except with $B_s = \{\theta : A_s(\theta,\alpha,n) \geq \sqrt{2}\}$ and changes in the integrand of (ii) as stated in the proposition. □

Proof of Theorem 2. For standard normal data we have $I(\theta) = \theta^2/2$ so that by Theorem 6

$$\mathbb{E}_{\theta\sim N(0,\tau)}E_\theta(T\wedge n) = nP_{\theta\sim N(0,\tau)}(\bar{A}) + \mathbb{E}_{\theta\sim N(0,\tau),A}(\mathbb{E}_\theta(T)).$$

Let $\delta = (2n^{-1}\log\alpha^{-1})^{1/2}$. The probability in the first term of the RHS is

$$P_{\theta\sim N(0,\tau)}(\bar{A}) = 2\int_0^\delta \frac{1}{\tau}\phi(\theta/\tau)d\theta \sim \frac{2\sqrt{2}}{\tau}\left(\frac{\log\alpha^{-1}}{n}\right)^{1/2}$$

by similar arguments to those in the proof of Proposition 1. For the second term, by equation (67) of [12] and the discussion following,

$$\begin{aligned}\mathbb{E}_\theta(T) &= 2\theta^{-2}\log\alpha^{-1} + \theta^{-2}\log\log\alpha^{-1} + D_1\theta^{-2}\log|\theta| \\ &\quad + D_2\theta^{-2} + D_3 + D_4\theta^{-1}B(\theta/2) + o(1)\end{aligned}$$

as $\alpha\to 0$, where

$$B(u) = \sum_{k=1}^\infty k^{-1/2}\phi(uk^{1/2}) - u\Phi(-uk^{1/2}).$$

The remainder of the proof requires verifying the order of the expectation of the above terms. First,

$$\begin{aligned}&\mathbb{E}_{\theta\sim N(0,\tau)}(\mathbf{1}_A I(\theta)^{-1}) \\ &= 4\int_\delta^\infty \theta^{-2}\frac{1}{\tau}\phi(\frac{\theta}{\tau})d\theta = \frac{4}{\tau^2}\left(\frac{\tau}{\delta}\phi(\frac{\delta}{\tau}) - \bar{\Phi}(\frac{\delta}{\tau})\right) < \infty\end{aligned}$$

for all $0 < \alpha < 1$ and $1 \neq n < \infty$, so that

$$\mathbb{E}_{\theta\sim N(0,\tau),A}2\theta^{-2}\log\alpha^{-1} \sim \frac{2\sqrt{2}}{\tau}n\left(\frac{\log\alpha^{-1}}{n}\right)^{1/2} = o(n).$$

By calculus,

$$\begin{aligned}&\mathbb{E}_{\theta\sim N(0,\tau),A}\theta^{-2}\log|\theta| \propto \int_\delta^\infty \theta^{-2}\log\theta e^{-\theta^2/2\tau^2} \\ &= \frac{1}{4}\left[\frac{1}{\sqrt{2}\tau}\Gamma\left(-\frac{1}{2}\,,\,\frac{\delta^2}{2\tau^2}\right)\log\delta\right. \\ &\left. + \delta^{-1}\text{MeijerG}\left(\{\{\},\{\frac{3}{2},\frac{3}{2}\},\{\{0,\frac{1}{2},\frac{1}{2}\},\{\}\},\frac{\delta^2}{2\tau^2}\right)\right],\end{aligned}$$

The MeijerG term is asymptotically constant as $\delta \to 0$, and

$$\Gamma\left(-\frac{1}{2}\,,\ \frac{\delta^2}{2\tau^2}\right) \to 2\sqrt{2}\tau\delta^{-1}.$$

It follows that

$$\begin{aligned}&\mathbb{E}_{\theta\sim N(0,\tau),A}\theta^{-2}\log|\theta| \\ &\propto n\left[K_1\left(n\log\alpha^{-1}\right)^{-1/2} + K_2\left(\frac{\delta\log\delta}{\log\alpha^{-1}}\right)\right]\end{aligned}$$

where K_i are both constants depending on τ. Both terms in the bracketed sum clearly converge to 0 since $\delta \to 0$. Next, we have by standard bounds of the normal CDF, $B(u) \geq 0$ and

$$\begin{aligned}B(u) &\leq u^{-2}\left(\int_1^\infty x^{-3/2}\phi(ux^{1/2})dx + \phi(u)\right) \\ &= \theta^{-2}\left(3\phi(\theta) - 2\theta\Phi(-\theta)\right)\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}_{\theta\sim N(0,\tau),A}\theta^{-1}B(\theta/2) &\leq K_3\mathbb{E}_{\theta\sim N(0,\tau),A}\theta^{-3}\phi(\theta/2) + K_4 \\ &\leq K_4\delta^{-2}e^{K_5\delta^2} - K_7\Gamma(0, K_6\delta^2) + K_4,\end{aligned}$$

where $\delta^{-2} = n/\log\alpha^{-2} = o(n)$ and $\Gamma(0, K_6\delta^2) \sim \log K_6/\delta^2 = O(\log\delta) = o(n)$.

Clearly the remaining terms are of lower order to those considered above. □

D Application to A/B Testing

The following lemma is a Central Limit Theorem approximation to the likelihood ratio. It justifies approximating Λ_n^H by $\tilde{\Lambda}_n^H$ when n is large.

Lemma 2. *For any* $0 < \bar{p} < 1, \theta \neq 0$

$$LR(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}; \bar{p}, \theta) = \frac{\phi_{(\theta,\hat{V}_{m,n})}(\bar{Y}_n - \bar{X}_m)}{\phi_{(0,\hat{V}_{m,n})}(\bar{Y}_n - \bar{X}_m)} + O(\min(m,n)^{-1/2})$$

where $\hat{V}_{m,n} = (\bar{X}_m(1-\bar{X}_m))/m + (\bar{Y}_n(1-\bar{Y}_n))/n$.

Proof. The pair $\bar{X}_m, \bar{Y}_n$ is sufficient for (p_0, p_1). Rotating both the statistics and the parameters, it follows that the pair $(\bar{X}_m + \bar{Y}_n), (\bar{Y}_n - \bar{X}_m)$ is sufficient for $(\bar{p}, \theta)$. By the CLT, up to order $\min(m,n)^{-1/2}$, the new pair of statistics are distributed as independent $N(2\bar{p}, V_n)$ and $N(\theta, V_n)$ respectively, where

$$V_n = \frac{p_0(1-p_0)}{m} + \frac{p_1(1-p_1)}{n} \tag{20}$$

α	n	Predicted Ratio	% average runtime for $r =$ 1e-4	1e-3	1e-2	0.1	1	10	Misspec. Error
0.001	10,000	0.047	0	0	36	64	0	0	2.2
0.010	10,000	0.039	0	0	40	56	4	0	2.5
0.100	10,000	0.027	0	0	52	40	8	0	4.6
0.001	100,000	0.015	0	24	76	0	0	0	4.3
0.010	100,000	0.012	0	48	56	0	0	0	4.6
0.100	100,000	0.009	0	48	48	4	0	0	7.4

Table 1: Simulation for optimal matching of mixing distribution over various choices of Type I error thresholds, α, and maximum sample size n. We define a τ^2 ratio as $r = \tau^2/\sigma^2$. The Predicted Ratio shows the τ^2 estimate from (8). The next 6 columns are the proportion of times the fixed r values were empirically found to have lowest average expected run time, $\mathbb{E}_G[T]$. This was done by sampling 200 experiment sample paths with effect sizes drawn from $G \sim N(0,1)$, and running a mSPRT with various rs. The r value with lowest average run time over sample paths was counted to have lowest expected run time for that replication. We perform $B = 25$ replications of each parameter combination. The most winning r values coincide well with the estimate in (8). The final column shows the average percent difference between the winning r's estimate of $\mathbb{E}_G[T]$ and the runner-up. A factor of 10 misspecification around the optimal r value results in a less than 10% increase in average runtime. Comparatively, there is over a factor of 2 difference between the winning and worst r estimates.

Hence

$$\mathsf{LR}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}); \bar{p}, \theta) \tag{21}$$

$$= \frac{\mathbb{P}_{\bar{p},\theta}(\bar{X}_m + \bar{Y}_n, \bar{Y}_n - \bar{X}_m)}{\mathbb{P}_{\bar{p},0}(\bar{X}_m + \bar{Y}_n, \bar{Y}_n - \bar{X}_m)} \tag{22}$$

$$= \frac{\phi_{(\theta,V_n)}(\bar{Y}_n - \bar{X}_m)}{\phi_{(0,V_n)}(\bar{Y}_n - \bar{X}_m)} + O(\min(m,n)^{-1/2}) \tag{23}$$

The result now follows because $\hat{V}_n \xrightarrow{p} V_n$ at rate $O(\min(m,n)^{-1})$. □

Proof of Theorem 4. Let $(\mathcal{F}_t)_{t=1}^{\infty}$ be the algebra generated by observations across the two streams in the order they arrive. The proof rests on the fact that Λ_t^H is a martingale under $(0, \bar{p})$ wrt this filtration. First we prove this in the case of an arbitrary deterministic allocation policy. Suppose wlog that the $(t+1)$th observation comes from stream $\mathbf{X}$. Then

$$\begin{aligned}\Lambda_{t+1}^H &= \int \mathsf{LR}(\mathbf{X}_{1:m(t)+1}, \mathbf{Y}_{1:n(t)}; \bar{p}, \theta)\, dH(\theta) \\ &= \int \mathsf{LR}(\mathbf{X}_{1:m(t)}, \mathbf{Y}_{1:n(t)}; \bar{p}, \theta) \left(\frac{\mathbb{P}_{\theta,\bar{p}}(X_{m(t)+1})}{\mathbb{P}_{0,\bar{p}}(X_{m(t)+1})}\right) dH(\theta)\end{aligned}$$

by independence. The martingale property holds because the term in parentheses is

independent of $\mathcal{F}_t$ and has expectation

$$\begin{aligned}&\mathbb{E}_{0,\bar{p}}\left(\frac{\mathbb{P}_{\theta,\bar{p}}(X_{m(t)+1})}{\mathbb{P}_{0,\bar{p}}(X_{m(t)+1})}\right)\\&=\int\left(\frac{\mathbb{P}_{\theta,\bar{p}}(x)}{\mathbb{P}_{0,\bar{p}}(x)}\right)\mathbb{P}_{0,\bar{p}}(x)dx=1\end{aligned}$$

Now, to generalize to arbitrary allocation policies, let A_t be the event that the tth observation belongs to $\mathbf{X}$. Then

$$\begin{aligned}&\Lambda_{t+1}^H\\&\quad=\int \mathsf{LR}(\mathbf{X}_{1:m(t)},\mathbf{Y}_{1:n(t)};\bar{p},\theta)L_{t+1}\,dH(\theta)\end{aligned}$$

where

$$L_{t+1}=1_{A_{t+1}}\left(\frac{\mathbb{P}_{\theta,\bar{p}}(X_{m(t)+1})}{\mathbb{P}_{0,\bar{p}}(X_{m(t)+1})}\right)+(1-1_{A_{t+1}})\left(\frac{\mathbb{P}_{\theta,\bar{p}}(Y_{n(t)+1})}{\mathbb{P}_{0,\bar{p}}(Y_{n(t)+1})}\right)$$

The key is that A_{t+1} is conditionally independent of $X_{m(t)+1}$ and $Y_{n(t)+1}$ given $\mathcal{F}_t$. Hence

$$\begin{aligned}\mathbb{E}_{\theta,\bar{p}}(L_{t+1}|\mathcal{F}_t)&=\mathbb{P}_{\theta,\bar{p}}(A_{t+1}|\mathcal{F}_t)\mathbb{E}_{0,\bar{p}}\left(\frac{\mathbb{P}_{\theta,\bar{p}}(X_{m(t)+1})}{\mathbb{P}_{0,\bar{p}}(X_{m(t)+1})}\right)+(1-\mathbb{P}_{\theta,\bar{p}}(A_{t+1}|\mathcal{F}_t))\,\mathbb{E}_{0,\bar{p}}\left(\frac{\mathbb{P}_{\theta,\bar{p}}(Y_{n(t)+1})}{\mathbb{P}_{0,\bar{p}}(Y_{n(t)+1})}\right)\\&=1\end{aligned}$$

This implies that Λ_t^H is a martingale as before.

Since $\Lambda_n^H\wedge\alpha^{-1}$ is then a bounded martingale, the Optional Stopping Theorem implies

$$1=\Lambda_0^H\wedge\alpha^{-1}=\mathbb{E}_{0,p_0}(\Lambda_{T(\alpha)}^H\wedge\alpha^{-1})\geq\alpha^{-1}\,P_{0,p_0}(T(\alpha)<\infty)\tag{24}$$

Rearranging gives the Type I error bound

$$P_{0,p_0}(\delta(\alpha)=1)=P_{0,p_0}(T(\alpha)<\infty)\leq\alpha\tag{25}$$

□

E Multiple Hypothesis Testing

q-values. For a given multiple testing procedure, the q-values are the vector $\mathbf{q}$, such that thresholding $\mathbf{q}$ at an arbitrary α obtains the same rejection set as the procedure applied to the p-values at that α. A q-value representation exists for any procedure where the rejection set is non-increasing in the p-values and non-decreasing in α. Clearly the minimum of two q-values is a q-value, so we will use the minimal representation. For Bonferroni, this is

$$q^i = (p^i m) \wedge 1.$$

For BH with independence or general dependence respectively,

$$q^{(j)} = \min_{k \geq j} \left(\frac{p^{(k)} m}{k} \right) \wedge 1 \text{ or } \min_{k \geq j} \left(\frac{p^{(k)} m \sum_{r=1}^{m} 1/r}{k} \right) \wedge 1.$$

FWER and FDR control: Proofs.

Proof of Proposition 2. $\forall \boldsymbol{\theta}$, the variables $p_T^1, ..., p_T^m$ satisfy the property that, for the truly null hypotheses i with $\theta_0^i = \theta^i$, p_T^i is marginally super-uniform. Hence there is a vector of (correlated) fixed-horizon p-values with the same distribution as $\mathbf{p}_T$, and so Bonferroni applied to the always valid p-values must control FWER. □

Before we can prove proposition 3, we require the following lemma.

Lemma 3.

$$\sup_{f \in \mathcal{F}} \sum_{k=1}^{m} \frac{1}{k} \int_{(k-1)\alpha/m}^{k\alpha/m} f(x)dx = \frac{\alpha}{m} \sum_{k=1}^{m} \frac{1}{k}$$

where $\mathcal{F} = \{f : [0,1] \to \mathbb{R}_+ \ : \ F(x) = \int_0^x f(x)dx \leq x \ , \ F(1) = 1\}$, $m \geq 1$, *and* $0 \leq \alpha \leq 1$.

Proof. Since $f \in \mathcal{F}$ are bounded, we restate the optimization in terms of $F_k = F(\frac{k\alpha}{m})$, and $F_0 \equiv 0$,

$$\begin{aligned} &\sup_{F_1,\ldots,F_m} \sum_{k=1}^{m} \frac{1}{k}(F_k - F_{k-1}) \\ &\text{subject to} 0 \leq F_j \leq \frac{k\alpha}{m} \ , \ F_k \geq F_{k-1} \ k = 1, \ldots, m. \end{aligned} \tag{26}$$

The objective can be rearranged as

$$\sum_{k=1}^{m-1} \frac{1}{k(k+1)} F_k + \frac{1}{m} F_m$$

which is clearly maximized by $F_k = \frac{k\alpha}{m}$ for all k. □

Proof of Proposition 3. As in the proof of Proposition 2, there is a vector of (correlated) fixed-horizon p-values with the same distribution as $\mathbf{p}_T$. However, we cannot simply invoke Theorem 1.3 in [3], which states that this form of BH controls FDR under arbitrary correlation, since that result requires that the fixed-horizon p-values be strictly uniform (rather than super-uniform). Nonetheless, adapting the proof of that theorem is straight-forward. Translating the proof into the sequential notation of this paper, the only non-immediate step is to show

$$\begin{aligned} &\sum_{k=1}^{m} \frac{1}{k} \sum_{r=k}^{m} \mathbb{P}\left(T_k^i \leq T < T_{k-1}^i \ , \ T_r \leq T < T_r^+ \ , \ T \leq \infty\right) \\ \leq &\sum_{k=1}^{m} \frac{1}{k} \mathbb{P}\left(\frac{(k-1)\alpha}{m} \leq p_T^i \leq \frac{k\alpha}{m} \right) \leq \frac{\alpha}{m} \sum_{k=1}^{m} \frac{1}{k} \end{aligned} \tag{27}$$

for all truly null hypotheses i. The first inequality is a restatement of definitions, and the second follows from Lemma 3 since by always-validity p^i_T is super-uniform. □

Proof of Theorem 5. We assume wlog that the truly null hypotheses are $i = 1, ..., m_0$. Letting V_n denote the number of true null rejected at n, the FDR can be expanded as

$$\begin{aligned}
&\mathbb{E}\left(\sum_{r=1}^{m} \frac{1}{r}\, V_T\, 1_{\{T_r \,\le\, T < T_r^+\}} 1_{T<\infty}\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{m_0}\sum_{r=1}^{m} \frac{1}{r}\, 1_{\{T_r^i \le T\}}\, 1_{\{T_r \,\le\, T < T_r^+\}} 1_{T<\infty}\right) \\
&= \sum_{i=1}^{m_0}\sum_{r=1}^{m} \frac{1}{r} \mathbb{P}\left(T_r^i \le T\ ,\ T_r \,\le\, T < T_r^+\ ,\ T<\infty\right)
\end{aligned}$$

Note that the sets $\{T_r \,\le\, T < T_r^+\}$ are disjoint and cover any location of T. Consider the terms in the sum over $i \in I$ and $i \notin I$ separately. For $i \notin I$, we bound the probability in the third equality by

$$\begin{aligned}
&\mathbb{P}\left(T_r^i \le T\ ,\ T<\infty\right) \mathbb{P}\left(T_r \le T < T_r^+ \,\middle|\, T_r^i \le T\ ,\ T<\infty\right) \\
&\le \frac{\alpha r}{M}\, \mathbb{P}\left(T_r \le T < T_r^+ \,\middle|\, T_r^i \le T\ ,\ T<\infty\right) \\
&= \frac{\alpha r}{M}\, \mathbb{P}\left((T_{r-1})_0^{-i} \,\le\, T < (T_{r-1})_0^{-i+} \,\middle|\, T_r^i \le T\ ,\ T<\infty\right)
\end{aligned}$$

where the first inequality follows from always-validity of sequential p-values, and the last equality because the modified BH procedure on the $m-1$ hypothesis other than the ith makes equivalent rejections at time T when $T_r^i \le T$.

For $i \in I$, arguing as in the proof of Proposition 3 shows

$$\begin{aligned}
&\sum_{r=1}^{m} \frac{1}{r} \mathbb{P}\left(T_r^i \le T\ ,\ T_r \,\le\, T < T_r^+\ ,\ T<\infty\right) \\
&\le \frac{\alpha}{m} \sum_{k=1}^{m} \frac{1}{k}.
\end{aligned}$$

The proof is completed on application of (13) to the terms in the first expansion with $i \notin I$ and re-ordering of the resulting terms. □

Examples. Here we apply Theorem 5 to three common classes of stopping times.

Example 1 (Rejection counts)**.** *Suppose we want some number $x \le m$ significant results to make a decision, and so stop at T_x. For each i*

$$\begin{aligned}
&\mathbb{P}\left((T_{r-1})_0^{-i} \,\le\, T_x < (T_{r-1}^+)_0^{-i} \,\middle|\, T_r^i \le T_x\ ,\ T_x<\infty\right) \\
&= \mathbb{P}\left(T_r \le T_x < T_r^+ \,\middle|\, T_r^i \le T_x\ ,\ T_x<\infty\right)
\end{aligned}$$

This probability is 1 if $r = x$ and 0 otherwise. Thus $I = \varnothing$, and so the FDR is at most $\alpha m_0/m$.

Example 2 (Rejecting a single hypothesis)**.**

$$T^2(k) = \inf\{t \ : \ R_t > 0 \ , \ p_t^k \leq qR_t/m\}.$$

be the first time hypothesis j is rejected. By independence, for any i,

$$\begin{aligned}&\mathbb{P}\left((T_{r-1})_0^{-i} \ \leq \ T < (T_{r-1}^+)_0^{-i} \ \middle| \ T_r^i \leq T \ , \ T < \infty\right)\\ &= \mathbb{P}\left((T_{r-1})_0^{-i} \ \leq \ T < (T_{r-1}^+)_0^{-i} \ \middle| \ T < \infty\right)\end{aligned}$$

Summing over r again we find that $I = \varnothing$. In fact, it is obvious that FDR should be controlled in this example since there is really only one hypothesis under consideration.

Example 3 (Rejecting a set of hypotheses)**.** *Let $T^{3\cap}(K)$ be the first time all hypotheses in some set $K \subseteq \{1, \ldots, m\}$ are rejected and $T^{3\cup}(K)$ be the first time any are rejected,*

$$T^{3\cap}(K) \doteq \max_{k\in K} T^2(k) \quad , \quad T^{3\cup}(K) \doteq \min_{k\in K} T^2(k).$$

Clearly, $T^{3\cap}(K)$ has equivalent bounds to the previous example. For $T^{3\cup}(K)$, we may argue that condition (13) holds for all $k \notin K$ as in the previous example. The condition does not hold for any $k \in K$, however, as knowing that $p_n^i \leq \alpha\frac{r}{m}$ at some n before the ith hypothesis is rejected makes it more likely that $R_{T^2(k)} \leq r$ than $R_{T^2(k)} > r$. Thus, we can only bound the FDR at

$$\frac{\alpha}{m}\left(m_0 + |\{1, \ldots, m_0\} \cap I| \sum_{k=2}^{m} \frac{1}{k}\right).$$

This bound is not tight though. In fact, when $m_0 = m$ and $K = \{1, \ldots, m_0\}$, we have $T^{3\cup}(K) = T_K$ so the FDR is at most α.

Confidence intervals and FCR. We focus on the case of independent data streams, with rejections made using BH-I. Firstly, Proposition 4 gives approximate FCR control in the fixed-horizon context, uniformly over the user's unknown selection rule.

Proof of Proposition 4. By Lemma 1 in [4],

$$FCR = \sum_{i=1}^{m}\sum_{r=1}^{m} \frac{1}{r}\mathbb{P}(|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin \tilde{CI}^i)$$

On the event $i \in J \cup S^{BH}$, there are two possibilities. If $i \in S^{BH}$, we can say $R^{BH} \leq |J \cup S_{BH}|$. If $i \notin S^{BH}$, we can say further that $R^{BH} + 1 \leq |J \cup S_{BH}|$. In either case, it follows that $CI^i(1 - \alpha|J \cup S_{BH}|/m) \subset \tilde{CI}^i$, and so the FCR is at most

$$\sum_{i=1}^{m}\sum_{r=1}^{m} \frac{1}{r}\mathbb{P}(|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin CI^i(1 - \alpha r/m))$$

Case 1: $i \notin J$.

$$\begin{aligned}
&\{|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin CI^i(1-\alpha r/m)\} \\
&\quad = \{|J \cup (S^{BH})_0^{-i}| = r-1, p^i \leq \alpha r/m, \theta^i \notin CI^i(1-\alpha r/m)\} \\
&\quad \subset \{|J \cup (S^{BH})_0^{-i}| = r-1, \theta^i \notin CI^i(1-\alpha r/m)\}
\end{aligned}$$

These two events are independent, so

$$\begin{aligned}
&\sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin CI^i(1-\alpha r/m)) \\
&\leq \sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup (S^{BH})_0^{-i}| = r-1) \mathbb{P}(\theta^i \notin CI^i(1-\alpha r/m)) \\
&\leq \frac{\alpha}{m} \sum_{r=1}^{m} \mathbb{P}(|J \cup (S^{BH})_0^{-i}| = r-1) = \frac{\alpha}{m}
\end{aligned}$$

Case 2: $i \in J$.

$$\begin{aligned}
&\sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin CI^i(1-\alpha r/m)) \\
&\leq \sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup S^{BH}| = r, \theta^i \notin CI^i(1-\alpha r/m)) \\
&= \sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup S^{BH}| = r \,|\, \theta^i \notin CI^i(1-\alpha r/m)) \mathbb{P}(\theta^i \notin CI^i(1-\alpha r/m)) \\
&\leq \frac{\alpha}{m} \sum_{r=1}^{m} \mathbb{P}(|J \cup S^{BH}| = r \,|\, \theta^i \notin CI^i(1-\alpha r/m))
\end{aligned}$$

Since S^{BH} is a function only of the p-values and the data streams are independent, the events $\{|J \cup S^{BH}| = r\}$ and $\{\theta^i \notin CI^i(1-\alpha r/m))\}$ are conditionally independent given p^i. Hence,

$$\mathbb{P}(|J \cup S^{BH}| = r \,|\, \theta^i \notin CI^i(1-\alpha r/m)) \leq \max_{\rho} \mathbb{P}(|J \cup S^{BH}| = r \,|\, p^i = \rho)$$

It is easily seen that this maximum must be attained at either $\rho = 0$ or $\rho = 1$, so

$$\begin{aligned}
&\mathbb{P}(|J \cup S^{BH}| = r \,|\, \theta^i \notin CI^i(1-\alpha r/m)) \\
&\qquad \leq \mathbb{P}(|J \cup S^{BH}| = r \,|\, p^i = 0) + \mathbb{P}(|J \cup S^{BH}| = r \,|\, p^i = 1) \\
&\qquad = \mathbb{P}(|J \cup (S^{BH})_0^{-i} \backslash i| = r-1) + \mathbb{P}(|J \cup (S^{BH})_1^{-i} \backslash i| = r-1)
\end{aligned}$$

Thus

$$\sum_{r=1}^{m} \frac{1}{r} \mathbb{P}(|J \cup S^{BH}| = r, i \in J \cup S^{BH}, \theta^i \notin CI^i(1 - \alpha r/m))$$
$$\leq \frac{\alpha}{m} \left\{ \sum_{r=1}^{m} \mathbb{P}(|J \cup (S^{BH})_0^{-i} \backslash i| = r - 1) + \sum_{r=1}^{m} \mathbb{P}(|J \cup (S^{BH})_1^{-i} \backslash i| = r - 1) \right\}$$
$$= \frac{2\alpha}{m}$$

Summing over all i now gives the desired result. □

Then this result carries over the sequential setting, if we make similar restrictions on the stopping time to those required for FDR control.

Definition 5. *If p_n^{i,θ_0} is the p-value for testing $H_0 : \theta^i = \theta_0$, let*

$$T_r^{i,\theta_0} = \inf\{t \ : \ p_t^{i,\theta_0} \leq \frac{\alpha r}{m}\} \tag{28}$$
$$(T_r)_0^{-i,J} = \inf\{t \ : \ |(S_n^{BH})_0^{-i} \cup J \backslash i| = r\} \tag{29}$$
$$(T_r^+)_0^{-i,J} = \inf\{t \ : \ |(S_n^{BH})_0^{-i} \cup J \backslash i| > r\}. \tag{30}$$

The last two stopping times denote the first times at least r and more than r experiments other than i are selected.

If $p_{(1),n}^{-i}, p_{(2),n}^{-i}, ...$ are the p-values for the experiments other than i placed in ascending order, consider another modified BH procedure that rejects hypotheses $(1), ..., (k)$ where k is maximal such that

$$p_{(k),n}^{-i} \leq \alpha \frac{k}{m},$$

These are the rejections obtained under the original BH-I procedure if $p_n^i = 1$. We define stopping times associated with this procedure $(T_r)_1^{-i,J}$ and $(T_r^+)_1^{-i,J}$ analogous to the two stopping times above.

Theorem 7. *Given independent always valid p-values $\mathbf{p}_n$ and corresponding CIs $(CI_n^i(1-s))_{i=1}^m$ at each level $s \in (0,1)$. Define new confidence intervals*

$$\tilde{CI}_n^i = \begin{cases} CI_n^i(1 - R_n^{BH}\alpha/m) & i \in S_n^{BH} \\ CI_n^i(1 - (R_n^{BH} + 1)\alpha/m) & i \notin S_n^{BH} \end{cases} \tag{31}$$

Let J be a set of experiments and let T be a stopping time such that the following conditions hold for every i, where θ^i is the true parameter value for that hypothesis:

$$\sum_{r=1}^{m} \mathbb{P}((T_r)_0^{-i,J} \leq T < (T_r^+)_0^{-i,J} | T_r^{i,\theta^i} \leq T < \infty) \leq 1 \tag{32}$$
$$\sum_{r=1}^{m} \mathbb{P}((T_r)_1^{-i,J} \leq T < (T_r^+)_1^{-i,J} | T_r^{i,\theta^i} \leq T < \infty) \leq 1 \tag{33}$$

Then under the selection rule $J \cup S_T^{BH}$, the intervals $(\tilde{CI}_T^i)$ have FCR at most $\alpha(1 + j/m)$.

Proof. By the same argument as in Proposition 4, we find that the FCR is at most

$$\sum_{i=1}^{m}\sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(|J \cup S_T^{BH}| = r, i \in J \cup S_T^{BH}, \theta^i \notin CI_T^i(1-\alpha r/m), T < \infty)$$

Case 1: $i \notin J$. As in Proposition 4, we obtain

$$\begin{aligned}
&\sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(|J \cup S_T^{BH}| = r, i \in J \cup S_T^{BH}, \theta^i \notin CI_T^i(1-\alpha r/m), T < \infty)\\
&\leq \sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(|J \cup (S_T^{BH})_0^{-i}| = r-1, \theta^i \notin CI_T^i(1-\alpha r/m), T < \infty)\\
&= \sum_{r=1}^{m}\frac{1}{r}\mathbb{P}((T_{r-1})_0^{-i,J} \leq T < (T_{r-1})_0^{-i,J+}, T_r^{i,\theta^i} \leq T < \infty)\\
&\leq \frac{\alpha}{m}\sum_{r=1}^{m}\mathbb{P}((T_{r-1})_0^{-i,J} \leq T < (T_{r-1})_0^{-i,J+} | T_r^{i,\theta^i} \leq T < \infty)\\
&\leq \frac{\alpha}{m}
\end{aligned}$$

Case 2: $i \in J$. As before,

$$\begin{aligned}
&\sum_{r=1}^{m}\frac{1}{r}\mathbb{P}(|J \cup S_T^{BH}| = r, i \in J \cup S_T^{BH}, \theta^i \notin CI_T^i(1-\alpha r/m))\\
&\quad\leq \frac{\alpha}{m}\sum_{r=1}^{m}\mathbb{P}(|J \cup S_T^{BH}| = r \,|\, \theta^i \notin CI_T^i(1-\alpha r/m))\\
&\quad= \frac{\alpha}{m}\sum_{r=1}^{m}\mathbb{P}(|J \cup S_T^{BH}| = r \,|\, T_r^{i,\theta^i} \leq T < \infty)\\
&\quad\leq \frac{\alpha}{m}\sum_{r=1}^{m}\max_{\rho}\mathbb{P}(|J \cup S_T^{BH}| = r \,|\, p_T^i = \rho, T_r^{i,\theta^i} \leq T < \infty)\\
&\quad\leq \frac{\alpha}{m}\left\{\sum_{r=1}^{m}\mathbb{P}(|J \cup (S_T^{BH})_0^{-i}\backslash i| = r-1 \,|\, T_r^{i,\theta^i} \leq T < \infty) + \sum_{r=1}^{m}\mathbb{P}(|J \cup (S_T^{BH})_1^{-i}\backslash i| = r-1 \,|\, T_r^{i,\theta^i} \leq T\right.\\
&\quad= \frac{\alpha}{m}\Big\{\sum_{r=1}^{m}\mathbb{P}((T_{r-1})_0^{-i,J} \leq T < (T_{r-1})_0^{-i,J+} \,|\, T_r^{i,\theta^i} \leq T < \infty)\\
&\qquad\qquad + \sum_{r=1}^{m}\mathbb{P}((T_{r-1})_1^{-i,J} \leq T < (T_{r-1})_1^{-i,J+} \,|\, T_r^{i,\theta^i} \leq T < \infty)\Big\}\\
&\quad\leq \frac{2\alpha}{m}
\end{aligned}$$

Finally we sum over i. □