Online Control of the False Discovery Rate under "Decision Deadlines"

Aaron Fisher
Foundation Medicine Inc.; 150 Second St, Cambridge, MA 02141

Abstract

Online testing procedures aim to control the extent of false discoveries over a sequence of hypothesis tests, allowing for the possibility that early-stage test results influence the choice of hypotheses to be tested in later stages. Typically, online methods assume that a permanent decision regarding the current test (reject or not reject) must be made before advancing to the next test. We instead assume that each hypothesis requires an immediate *preliminary* decision, but also allows us to update that decision until a preset deadline. Roughly speaking, this lets us apply a Benjamini-Hochberg-type procedure over a moving window of hypotheses, where the threshold parameters for upcoming tests can be determined based on preliminary results. We show that our approach can control the false discovery rate (FDR) at every stage of testing, even under arbitrary p-value dependencies. That said, our approach offers much greater flexibility if the p-values exhibit a known independence structure. For example, if the p-value sequence is finite and all p-values are independent, then we can additionally control FDR at adaptively chosen stopping times.

Keywords: adaptive stopping time, batch testing, data decay, decaying memory, quality preserving database.

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Scientific discoveries form an ongoing, ever-evolving process. Each new experiment offers an opportunity to suggest new hypotheses based on results that have come before. Traditionally, the hypotheses researchers plan to test in an experiment are prespecified before any data from the experiment is visible, as this facilitates control of either the false discovery rate (FDR; Benjamini and Hochberg, 1995) or the probability of producing *any* false positives (the familywise error rate, or FWER; see, for example Efron and Hastie, 2016) within that experiment.

In contrast to fully prespecified procedures, *online* procedures test hypotheses sequentially, and allow the results of preliminary tests to inform choices about which hypotheses to focus on in future tests (Foster and Stine, 2008). These procedures typically require that error rates be controlled at every stage of the sequence (e.g., Javanmard and Montanari, 2015; Ramdas et al., 2017). The online setting is increasingly relevant to large-scale experimentation, and to repeated analyses of public datasets (Aharoni and Rosset, 2014). At a high level, online testing can be seen as an abstraction of the scientific process itself (Xu and Ramdas, 2020).

Online testing problems also arise when users must quickly decide how to take action in response to a stream of data. Applications range from monitoring credit card transactions for instances of fraud (Zrnic et al., 2020) to deciding how to assign treatments to sequences of patients. Here, hypotheses quickly become irrelevant, and so final decisions must be made without delay. In other words, a discovery has little value if the opportunity to act on it has passed.

On the other hand, streams of hypothesis tests do not always require immediate, permanent decisions. In particular, if our goal is to maintain a growing library of scientific knowledge (Aharoni and Rosset, 2014), then hypotheses can remain relevant long after they are tested. Here, discoveries remain valuable even if they are made *retroactively*.

With this mind, we study scenarios where limited

forms of decision updating still add value. Specifically, we consider the setting where each hypothesis requires an immediate, preliminary decision (reject or not reject), but also allows us to update that decision until some preset deadline. To incorporate these "decision deadlines," we blend two existing procedures: the well-known, offline Benjamini and Hochberg (BH, 1995) procedure, and an online procedure known as *significance levels based on number of discoveries* (LOND; Javanmard and Montanari 2015; see also Zrnic et al., 2021). Our procedure can reduce to LOND if all decisions must be made immediately, or to BH if all decisions can be updated indefinitely. Because the option for decision updates is limited to evolving subset of "active" hypotheses, we refer to our approach as *significance thresholds based on active discoveries* (TOAD).

We show that our approach can provide online FDR control under arbitrary p-value dependencies. That said, our approach offers greater power and flexibility when the p-values follow a known independence structure (e.g., independence across batches). In such cases, we allow the parameters used in setting significance thresholds to be determined based on preliminary results. This, in turn, lets us employ certain types of adaptive stopping rules without sacrificing FDR control. As a simple example, if our p-value sequence has finite length and follows independence, then we can still control FDR even if analysts end their experiments at any time due to especially strong preliminary results.

1.1 Outline

The remained of our paper is organized as follows. Section 1.2 discusses the advantages of our approach relative to other methods in the literature. Section 1.3 introduces relevant notation. Section 2 presents the TOAD procedure along with its FDR guarantees. Section 3 uses simulations to compare the power of TOAD to the power of similar methods introduced by Zrnic et al. (2020). We conclude with a discussion of how our approach could be extended to incorporate the concept of "decaying memory" (Ramdas et al., 2017).

All proofs are provided in the appendix. These proofs use a combination of methods from Blanchard and Roquain, 2008; Javanmard and Montanari, 2015; Ramdas et al., 2017; and Zrnic et al., 2021. The appendix also includes additional simulations, and an applied analysis of fraud detection in credit card transactions (as in Zrnic et al., 2020).

1.2 Related Literature

In recent work that most closely resembles our own, Zrnic et al. (2020) propose two online methods for applying Benjamini-Hochberg procedures to *batches* of hypotheses (referred to as Batch_{BH} and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$). This batch testing framework forms a special case of online testing under decision deadlines, where the deadline for each test in a batch is the time of the last test in that batch.

Our work differs from that of Zrnic et al. (2020) in three substantial ways. First, our framing in terms of "deadlines" is more flexible than the batch structure used by Zrnic et al.. Second, we will show analytically that TOAD is at least as powerful as $\text{Batch}_{\text{BH}}^{\text{PRDS}}$, and will show in simulations that it is typically more powerful (see Sections 2 & 3, as well as the appendix). Finally, certain versions of our approach control FDR under arbitrary p-value dependencies, whereas Zrnic et al. (2020) prove FDR control under an assumption of independence across batches.

In another approach that is conceptually similar to ours, Zrnic et al. (2021) suggest "revisiting" hypotheses by allowing duplicated test statistics in later stages (see their Section 3). We differ from Zrnic et al. (2021) in that we simultaneously update all active hypotheses at every stage rather than updating hypotheses individually.

The fact that TOAD can provide online FDR control under arbitrary p-value dependencies is nontrivial in the literature. Typically, online bounds on the FDR require either an independence condition on the p-values (Ramdas et al., 2017, 2018; Tian and Ramdas, 2019; Zrnic et al., 2020, 2021), or a positive dependence condition (Fisher, 2021). Three notable exceptions include the (reshaped) LOND method (Javanmard and Montanari, 2015; Zrnic et al., 2021), upon which TOAD is based; certain forms of generalized alpha investing rules (Javanmard and Montanari, 2018, see their Theorem 3.6); and the SupLORD method (Xu and Ramdas, 2020), which controls FDR under a dependence condition known as *conditional superuniformity*. We will employ this same condition in the next section, in order to improve the flexibility of TOAD. As an alternative to the traditional FDR, many online testing methods instead focus on controlling either the "modified" FDR or the marginal FDR (Foster and Stine 2008; Aharoni and Rosset 2014; Ramdas et al. 2017, 2018; Tian and Ramdas 2019; Zrnic et al. 2021).

1.3 Notation

Let $H_1, H_2 \ldots$ be a possibly infinite sequence of hypotheses, and let $P_1, P_2, \ldots$ be p-values associated with each hypothesis. Such a sequence can result either from a growing (streaming) dataset with an increasing number of subgroups, or from a series of dis-

tinct questions applied to a fixed dataset. As we discuss in the appendix, many forms of online decision making can be captured by this framework.

We consider the setting where, at each stage t of testing, we observe the next p-value P_t and must make an immediate, preliminary decision to reject or not reject H_t. However, we are also permitted to update our decision up until a preset deadline $d_t \geq t$ (i.e., the decision for H_t cannot be altered after stage d_t). We use $\mathcal{C}_t$ to denote the set of "active" candidate hypotheses for which decisions can still be updated at stage t, i.e., $\mathcal{C}_t = \{i \leq t : d_i \geq t\}$. For example, if we allow rejection decisions to be updated indefinitely, then $d_t = \infty$ and $\mathcal{C}_t = \{1, \ldots, t\}$ for all t. If we require final decisions instantaneously, then $\mathcal{C}_t = \{d_t\} = \{t\}$.

Let $\mathcal{R}_t \subseteq \{1, \ldots, t\}$ denote the indices for the hypotheses that we reject at stage t. Again, any differences in the sets of hypotheses rejected at consecutive stages must be limited to the hypotheses whose deadlines have not yet passed (i.e., $\{\mathcal{R}_t \setminus \mathcal{C}_t\} = \{\mathcal{R}_{t-1} \setminus \mathcal{C}_t\}$).

We define $\mathcal{H}_0 \subseteq \mathbb{N}$ to be the indices corresponding to true null hypotheses, and define the FDR at time t to be

$$\text{FDR}(t) = \mathbb{E}\left[\frac{|\mathcal{H}_0 \cap \mathcal{R}_t|}{1 \vee |\mathcal{R}_t|}\right],$$

where $a \vee b$ denotes the maximum over $\{a, b\}$. We use α to denote a desired level at which to control $\text{FDR}(t)$.

2 THRESHOLDS BASED ON ACTIVE DISCOVERIES (TOAD)

We first describe the original LOND procedure (Javanmard and Montanari, 2015), as this method forms the original inspiration for our proposed method. As input, LOND requires a sequence of nonnegative tuning parameters $a_1, a_2, \ldots$ satisfying $\sum_{i=1}^{\infty} a_i = 1$. At each stage t, LOND rejects H_t if

$$P_t \leq (|\mathcal{R}_{t-1}| + 1) a_t \alpha. \tag{1}$$

Once a hypothesis is rejected, it remains rejected in all future stages. Javanmard and Montanari (2015) show that, under a condition on the joint distribution of p-values, LOND controls FDR at every stage.

Building on this method, Zrnic et al. (2021) propose a "reshaped" version of LOND that controls FDR under any p-value dependency structure (see also Theorem 2.7 of Javanmard and Montanari, 2015). This version additionally takes as input a sequence of so-called *shape functions* $\{\beta_i\}_{i=1}^{\infty}$. Following Blanchard and Roquain (2008), we say that β is a shape function if there exists a probability distribution ν on $\mathbb{R}_{>0}$ such that

$$\beta(r) = \mathbb{E}_{X \sim \nu}\left[X \times 1(X \leq r)\right]. \tag{2}$$

For example, when the number of stages ($t_{\max}$) is finite, Blanchard and Roquain consider setting ν to be the distribution satisfying $\mathbb{P}_{X \sim \nu}(X = x) \propto 1/x$ for each $x \in \{1, \ldots, t_{\max}\}$. This produces the shape function $\beta(r) = r\left(\sum_{i'=1}^{t_{\max}} 1/i'\right)^{-1}$, which mimics the transformation employed by Benjamini and Yekutieli (2001). To incorporate these shape functions $\{\beta_i\}_{i=1}^{\infty}$, Zrnic et al. define the reshaped version of LOND to reject each H_t whenever $P_t \leq \beta_t(|\mathcal{R}_{t-1}| \vee 1) a_t \alpha$.

Our proposed procedure differs from (reshaped) LOND in three key ways. The first is a restriction, which is that we require users to select a common function β to be used at all stages. More specifically, users can set β to be either the identity function or a shape function. Setting β to be the identity function is the simplest and most powerful option, but setting β to be a shape function will improve our FDR guarantee (see details in Section 2.1).

The second two differences are expansions. Rather than prespecifying all parameters $\{a_i\}_{i=1}^{\infty}$, we replace them with random nonnegative random variables $\{A_i\}_{i=1}^{\infty}$ satisfying $\sum_{i=1}^{\infty} A_i = 1$, where each A_i can depend on the previously observed p-values. We define $\tau_i \leq i - 1$ to be the stage by which the i^{th} parameter A_i must be selected. That is, we require A_i to be a deterministic function of the first τ_i p-values $\{P_{i'}\}_{i' \leq \tau_i}$.

In practice, we will see in the next section that our ability to choose τ_i while still controlling FDR will depend on our knowledge of the p-value dependence structure. For example, if the p-values are known to be independent, we can set $\tau_i = i - 1$. However, if no assumptions can be confidently made of the p-value dependence structure, then we are typically limited to setting each $\tau_i = 0$, as in the original LOND procedure.

We also expand on LOND by allowing users to update rejection decisions for hypotheses whose deadlines have not yet passed. At each stage t, our goal will be to find the *largest set* of rejected indices $\mathcal{R}_t \subseteq \{1, \ldots, t\}$ that satisfies the following two properties: (1) decisions for nonactive hypotheses are not updated ($\{\mathcal{R}_t \setminus \mathcal{C}_t\} = \{\mathcal{R}_{t-1} \setminus \mathcal{C}_t\}$), and (2) for all $i \in \mathcal{R}_t$, we have $P_i \leq \beta(1 \vee |\mathcal{R}_t|) A_i \alpha$. The second property mimics the LOND condition (Eq (1)), and will be used to show FDR control. We achieve these two properties as follows.

Algorithm 1. (TOAD) Take as input a function β (either the identity function or a shape function), and a value for A_1.

1. (Initialize) Set $\mathcal{R}_0 = \emptyset$. For any $i \in \mathbb{N}$ such that $\tau_i = 0$, determine the value for A_i.
2. For each stage t:

(a) (Save past rejections) Define $\mathcal{R}_t^{\text{old}} = \mathcal{R}_{t-1} \backslash \mathcal{C}_t$ to be the set of previously rejected indices that are no longer being actively updated.

(b) (Order test statistics) Let $W_i = P_i/A_i$, and let $W_{(j,t)}$ be the j^{th} lowest value from the set $\{W_i\}_{i \in \mathcal{C}_t}$, such that $W_{(1,t)} \leq \cdots \leq W_{(|\mathcal{C}_t|,t)}$.

(c) (Define current rejections) Reject the set of indices $\mathcal{R}_t = \mathcal{R}_t^{\text{old}} \cup \{i \in \mathcal{C}_t : W_i \leq W_{(S_t,t)}\}$, where

$$S_t = \max\{j \leq |\mathcal{C}_t| : W_{(j,t)} \leq \alpha\beta(j + |\mathcal{R}_t^{\text{old}}|)\}. \quad (3)$$

(d) (Set threshold parameters) For any $i > t$ such that $\tau_i = t$, determine the value for A_i.

While TOAD can retroactively *reject* certain hypotheses, we show in the appendix that TOAD never *reverses* a previous rejection (i.e., $\mathcal{R}_t \subseteq \mathcal{R}_{t'}$ for any $t < t'$). This monotonicity property is not strictly required by our framing, but may facilitate the procedure's implementation. For example, the property can prove useful if it is logistically straightforward to announce a new discovery, but difficult to retract a previously announced discovery.

We can think of TOAD as a generalization of both (reshaped) LOND and BH. In the special case where all rejection decisions must be finalized immediately (i.e., $\mathcal{C}_t = \{t\}$) and β is the identity function, our procedure reduces to a version of LOND with dynamically defined threshold parameters. If $\mathcal{C}_t = \{t\}$ and β is a shape function, then TOAD recovers the reshaped LOND method studied by Zrnic et al. (2021). At the other extreme, if our hypothesis sequence contains a finite number of elements (denoted by $t_{\max}$), and if all hypotheses remain active for the entire sequence (i.e., $\mathcal{C}_{t_{\max}} = \{1, \ldots, t_{\max}\}$), then we can recover the BH algorithm setting $A_i = 1/t_{\max}$ for all i, setting β to be the identity function, and applying TOAD at stage $t_{\max}$.

As an intermediate setting, if hypotheses remain active according to a block structure then we can recover a procedure that closely resembles the Batch_{BH} and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ algorithms described by Zrnic et al. (2020). In fact, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ can *also* be seen as a generalization of both BH and LOND (Zrnic et al., 2020).

However, we show in the appendix that any hypothesis rejected by $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ is also rejected by TOAD. Our simulations in Section 3 show that the reverse is not true, and that TOAD typically achieves higher power than $\text{Batch}_{\text{BH}}^{\text{PRDS}}$. One approximate explanation for this power difference is that TOAD makes more modifications to the underlying BH procedure than $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ does. $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ can be implemented by repeatedly applying BH to a sequence of batches, using the number of previous rejections to inform the *alpha-level* used for the next application of BH. By contrast, TOAD modifies the BH procedure itself by adding the number of previous rejections ($|\mathcal{R}_t^{old}|$) in Eq (3). More formally, for β equal to the identity function, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ can be seen as replacing $\alpha\beta(j + |\mathcal{R}_t^{\text{old}}|) = \alpha \times (j + |\mathcal{R}_t^{\text{old}}|)$ in our Eq (3) with the smaller quantity $\left\{\alpha\left(1 + |\mathcal{R}_t^{\text{old}}|/|\mathcal{C}_t|\right)\right\} \times j$ (see details in the appendix).

2.1 FDR Control

Next, we outline sufficient conditions for FDR control. Our first assumption places restrictions on how the thresholds can be selected. This assumption can be ensured by design.

Assumption 1. (Threshold selection) For each $i \in \mathbb{N}$, A_i is a deterministic function of the first τ_i p-values, denoted by $\mathcal{P}_{\tau_i} = \{P_1, \ldots, P_{\tau_i}\}$.

Next, we assume that users have access to conditionally valid test statistics for each hypothesis. Specifically, we assume that the p-value P_i for any true null H_i is conditionally (super)uniformly distributed, given the information used to select A_i.

Assumption 2. (Conditional super-uniformity) For any $i \in \mathcal{H}_0$, we have $\mathbb{P}(P_i \leq u | \mathcal{P}_{\tau_i}) \leq u$ for all $u \in [0, 1]$ and all realizations of $\mathcal{P}_{\tau_i}$.

This assumption is based on super-uniformity assumptions used by Foster and Stine (2008, see their Eq (10)); Aharoni and Rosset (2014, see their Assumption 1); Ramdas et al. (2017); Xu and Ramdas (2020) and Zrnic et al. (2021). The assumption is also conceptually similar to a condition used by Javanmard and Montanari (2015, see their Eq (8)).

In practice, Assumption 2 can often restrict us to define parameters A_i only based on the previous p-values that are *independent* of P_i, that is, setting τ_i far enough in advance to ensure that $\{P_{i'}\}_{i' \leq \tau_i} \perp P_i$. Zrnic et al. (2021) point out that such a strategy is sufficient to ensure Assumption 2. Moreover, Fisher (2021) show that if P_i is continuous and well-calibrated, then the *only* way to satisfy Assumption 2 is if $\{P_{i'}\}_{i' \leq \tau_i} \perp P_i$. Here, the term well-calibrated means that there exists no alternative, more powerful p-value P_i' that can be defined as a deterministic function of $\{P_1, \ldots, P_{\tau_i}, P_i\}$ such that P_i' also satisfies Assumption 2, $P_i' \leq P_i$, and $\mathbb{P}(P_i' \neq P_i) > 0$. Zrnic et al. also note that setting parameters in advance is a natural way to capture the logistical delays that can occur between test specification and test completion.

Next, we define a condition regarding positive dependence of the p-values.

Assumption 3. (Conditional positive dependence) For any set of positive integers $\{t, r, i\}$ satisfying $r, i \leq t$

and $H_i \in \mathcal{H}_0$, the probability

$$\mathbb{P}(1 \vee |\mathcal{R}_t| \leq r | P_i \leq u, \mathcal{P}_{\tau_i})$$

is nondecreasing in u.

Roughly speaking, Assumption 3 says that higher p-values imply a higher probability that $|\mathcal{R}_t|$ is small. We show in the appendix that this condition holds under a version of "positive regression dependence on a subset" (PRDS; Benjamini and Yekutieli, 2001), as well as a monotonicity assumption applied to the parameters A_i. This monotonicity assumption is satisfied if, for example, any adaptively defined parameter A_i is monotonically decreasing in the p-values observed so far.

We are now prepared to show FDR control for our procedure.

Theorem 1. *(FDR Control) Under Assumptions 1 & 2, TOAD satisfies FDR$(t) \leq \alpha$ for any $t \in \mathbb{N}$ if either of the following conditions hold:*

1. *(Positive dependence) Assumption 3 holds and β is the identity function; or*
2. *(General dependence) β is a shape functions in the form of Eq (2).*

Theorem 1 carries immediate implications for controlling $\mathbb{E}[FDR(T)]$ at an adaptively determined stopping time T. If our hypothesis sequence is finite, and if we can define our parameters A_t adaptively while still satisfying Assumption 2 (see discussion above), then we can incorporate an adaptive stopping time T by simply setting $A_t = 0$ for all $t > T$, and completing the test procedure up to and including the final stage.

That said, there are two important caveats to this way of capturing adaptive stopping times. The first is that certain adaptive stopping rules may lead to violations of Assumption 3, requiring us to either carefully verify this assumption or to appeal to Part 2 of Theorem 1 instead. The second is that these forms of adaptive stopping rules become limited when researchers set parameters A_i several stages in advance ($\tau_i < i - 1$). By specifying the parameter for a future test, a researcher also implicitly commits to *completing* that future test. Although they can adaptively choose to stop all testing for stages where parameters have not yet been determined, they cannot choose to avoid tests that have already been specified.

3 SIMULATIONS

In this section, we investigate the effect of the deadline structure on TOAD's power. We also compare TOAD against two methods introduced by Zrnic et al. (2020), and against a "naive" version of BH.

We adopt a simulation setup based the one used by Zrnic et al. (2020; differences are noted below). We define a sequence of $t_{\max} = 3000$ test statistics $(Z_1, \dots Z_{t_{\max}}) \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, \dots, \mu_{t_{\max}})$ is a sequence of mean parameters and Σ is a covariance matrix defined in detail below. For each test statistic Z_i, our null hypothesis H_i is that $E(Z_i) = 0$, and our alternative hypothesis is that $E(Z_i) = 3$. We use π_1 to denote the proportion of null hypotheses that are false. In each simulation iteration, we select a random subset of $\lceil (1-\pi_1) t_{\max} \rceil$ indices for which we set $\mu_i = 0$ (i.e., we simulate Z_i from the null distribution). We set the remaining mean parameters equal to 3.

To define deadline parameters, we will say that hypotheses remain active within "batches" of tests, and use n_{batch} to denote the batch size. For each $i \in \{1, \dots, t_{\max}\}$, we set the deadline d_i to be the smallest multiple of n_{batch} that is no less than i, that is, $d_i = \min\{k n_{\text{batch}} : k \in \mathbb{N} \text{ and } i \leq k n_{\text{batch}}\}$. For example, if $n_{\text{batch}} = 100$, then $d_i = 100$ for $i \in [1, 100]$; $d_i = 200$ for $i \in [101, 200]$; and so on. We define Σ so that $Var(Z_i) = 1$ for all i; $Cov(Z_i, Z_j) = \rho$ if $i \neq j$, but i and j are in the same batch; and $Cov(Z_i, Z_j) = 0$ if i and j are not in the same batch

We simulate all combinations of $\rho \in \{0, 0.5\}$; $n_{\text{batch}} \in \{10, 100, 1000\}$; and

$$\pi_1 \in \{0.01, 0.02, \dots, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5\}.$$

For each combination, we simulate 500 iterations.

Our simulation setup differs from that of Zrnic et al. (2020) in two ways. Most notably, Zrnic et al. only simulate the case where $\rho = 0$, as most of the methods they develop are designed for the case of independent test statistics. Zrnic et al. also use a Bernoulli distribution to determine whether each test statistic Z_i is generated from a null distribution or an alternative distribution, meaning that the realized proportion of truly null hypotheses varies slightly across simulation iterations.

3.1 Comparator Methods

As comparators for TOAD, we primarily consider the Batch_{BH} and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ algorithms (Zrnic et al., 2020). The first method, Batch_{BH}, is proven to control FDR under an independence assumption. The second method, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$, is proven to control FDR if test statistics are independent across batches and positively dependent within each batch. Thus, we expect Batch_{BH} to achieve higher power than $\text{Batch}_{\text{BH}}^{\text{PRDS}}$, potentially at the cost of FDR control.

For the tuning parameters of TOAD, we set β equal to the identity function, and set $\tau_i = 0$ and $A_i = 1/t_{\max}$ for all i. Similarly, for $\text{Batch}_{\text{BH}}^{\text{PRDS}}$, we use the implementation defined in Zrnic et al.'s appendix, and use tuning parameters that place equal weight on each batch. For Batch_{BH}, we use the implementation and tuning parameters described in Zrnic et al.'s simulations.

We also compare against the "naive" approach of running BH separately in each batch at an alpha level of $\alpha(t_{\max}/n_{\text{batch}})^{-1}$, where $t_{\max}/n_{\text{batch}}$ is the number of batches. We refer to this last method as "Naive-BH." For completeness, we briefly show in the appendix that Naive-BH also controls the false discovery rate whenever the p-values are positively dependent.

For all of the above methods, we set $P_i = \Phi(-Z_i)$, where Φ is the CDF of a standard normal distribution. That is, we define each p-value to be the result of a one-sided test of H_i.

3.2 Simulation Results

Figure 1 shows the simulated power for each method tested, where power is defined as the expected proportion of alternative hypotheses that are rejected in any one experiment. Figure 2 shows the FDR for each procedure.

Batch_{BH} consistently generates the highest power, with TOAD generating the second highest. The one exception comes when batches sizes are large ($b = 1000$), in which case TOAD and Batch_{BH} have comparable power. To some extent, this is to be expected, as TOAD provides stronger FDR guarantees than Batch_{BH} does. Indeed, we see that when the assumptions of Batch_{BH} are violated due to within-batch correlation, Batch_{BH} produces an inflated FDR (see Figure 2).

On the other hand, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ offers FDR guarantees that are more comparable to those of TOAD. Thus, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ forms an especially informative comparator. We see that TOAD has higher power than $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ across all scenarios, as we would expect from our analytical result in the appendix.

In addition to these experiments, we simulated cases where the test statistics follow a first order autoregressive process, as well as cases where each mean parameter μ_i corresponding to an alternative distribution is randomly generated (as in Zrnic et al., 2020). Neither variation led to a substantial change in results (see details in the appendix).

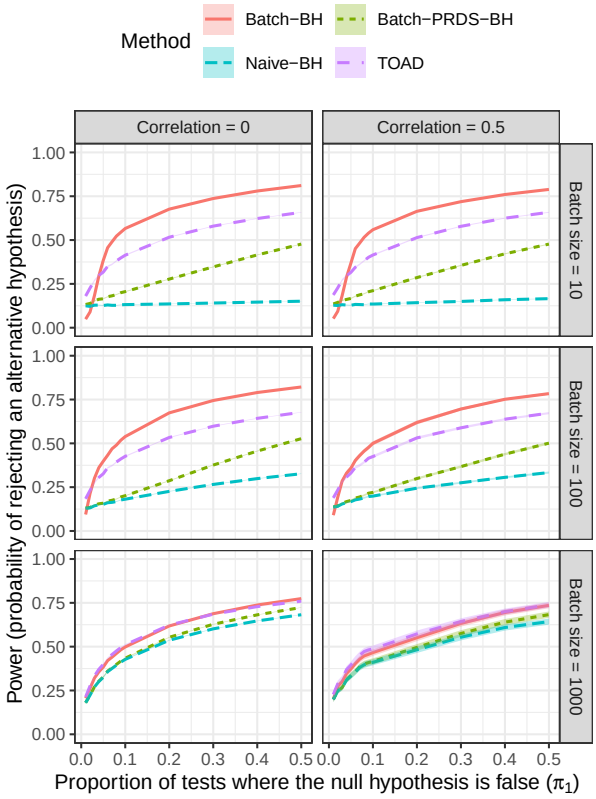

Figure 1: Simulated power for each method – We simulate test statistics under a "batch" structure, where all hypotheses in a batch share a common deadline. The test statistics are normally distributed with possible within-batch correlation (denoted by columns). For each null hypothesis $H_i : \mathbb{E}(Z_i) = 0$, we generate one-sided p-values as $\Phi(-Z_i)$, where Φ is the cumulative distribution function for a standard normal distribution. Shaded ribbons show a range of $\pm$ two Monte Carlo standard errors ($\sqrt{\frac{1}{500}\text{Var}\left(|\mathcal{R}_{t_{\max}} \cap \bar{\mathcal{H}}_0|/|\bar{\mathcal{H}}_0|\right)}$, where 500 is the number of simulation iterations and $\bar{\mathcal{H}}_0$ is the set of false nulls), although these errors are negligible in many cases. The Batch_{BH} method generates the highest power, but also requires the strongest assumptions in order to guarantee control of the FDR. Of the methods that ensure FDR control for positively dependent test statistics, TOAD achieves the highest power.

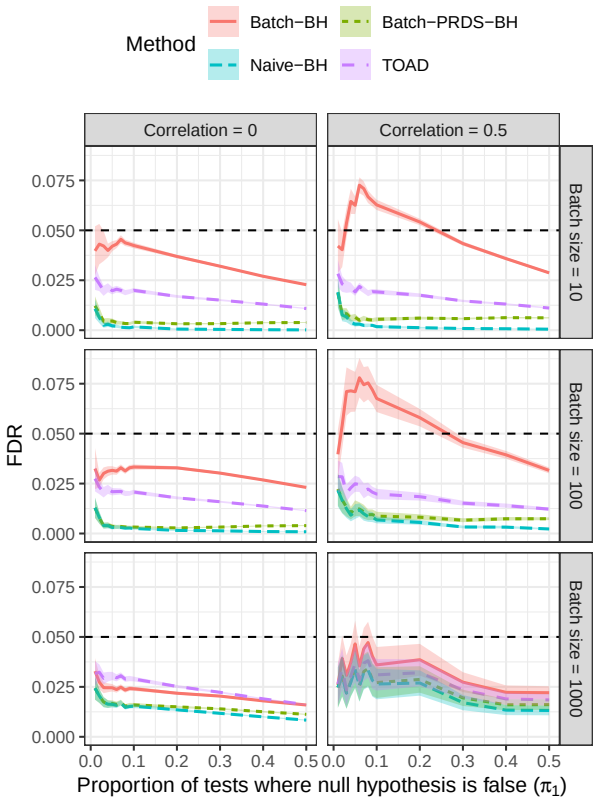

Figure 2: Simulated FDR for each method – Again, shaded ribbons show a range of $\pm$ two Monte Carlo standard errors ($\sqrt{\frac{1}{500}\text{Var}\left(|\mathcal{H}_0 \cap \mathcal{R}_{t_{\max}}| / \left(1 \vee |\mathcal{R}_{t_{\max}}|\right)\right)}$, where 500 is the number of simulation iterations). The dashed line shows our desired FDR level. We see that the power of Batch_{BH} can come at the cost of inflated FDR in the face of within-batch correlation (right column).

4 DISCUSSION

We have proposed an online version of the Benjamini and Hochberg (1995) method that allows for limited forms of decision updating. Our procedure controls the FDR under arbitrary p-value dependence structures. If partial knowledge of the p-value dependence structure is available, we can additionally control FDR under certain forms of adaptive stopping rules. Compared to similar procedures with comparable FDR guarantees, we find that our approach also provides superior power.

We conclude by discussing an immediate extension that incorporates the concept of "decaying memory." Ramdas et al. (2017) remark that, in short-term forecasting problems, hypotheses tested in the distant past have little bearing on our decisions at present. With this in mind, they propose a "decaying memory" variation of FDR that places more weight on recently tested hypotheses. That is, they focus on multiplicity corrections for the discoveries currently in use, rather than for all discoveries made over the course of an experiment.

In some ways, the idea that hypotheses from the distant past carry less importance at present is a natural complement to the idea that hypotheses eventually pass a deadline beyond which any retroactive discovery is irrelevant. Thus, one fruitful avenue of future research could be to formally blend the ideas of decaying memory and deadlines.

A simply way of doing this is to omit "outdated" or "forgotten" hypotheses from the FDR computation, resulting in

$$\text{FDR}_{\text{recent}}(t) = \mathbb{E}\left[\frac{|\mathcal{H}_0 \cap \mathcal{R}_t \cap \mathcal{C}_t|}{1 \vee |\mathcal{R}_t \cap \mathcal{C}_t|}\right].$$

It is straightforward to show that TOAD controls $\text{FDR}_{\text{recent}}(t)$ if we (1) relax the requirement that $\sum_{i=1}^{\infty} A_i \leq 1$ to instead require that $\sum_{i \in \mathcal{C}_t} A_i \leq 1$ for all t, and (2) replace $\mathcal{R}_t^{\text{old}}$ with the empty set $\emptyset$ throughout the procedure (see the appendix). Under such a procedure, the parameters A_i from outdated hypotheses can be "recycled" towards future tests.

However, an important caveat is that $\text{FDR}_{\text{recent}}(T)$ is more difficult to control under adaptive stopping times T. Before, we were able to control $\text{FDR}(T)$ simply by controlling $\text{FDR}(t_{\max})$ (Section 2.1). Here though, controlling $\text{FDR}_{\text{recent}}(t_{\max})$ is not sufficient for controlling $\text{FDR}_{\text{recent}}(T)$. Roughly speaking, $\text{FDR}_{\text{recent}}(t_{\max})$ "forgets" the information that would have been necessary to control error rates at earlier times.

Acknowledgements

I am grateful for a helpful correspondence with Tijana Zrnic during the early phases of this work.

References

Aharoni, E. and Rosset, S. (2014). Generalized α-investing: definitions, optimality results and application to public databases. *J. R. Stat. Soc. Series B Stat. Methodol.*, 76(4):771–794.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188.

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *EJSS*, 2(none):963–992.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Fisher, A. (2021). Online false discovery rate control for LORD & SAFFRON under positive, local dependence. *arXiv [stat.ME]*.

Foster, D. P. and Stine, R. A. (2008). $\mathcal{A}$-investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Series B Stat. Methodol.*, 70(2):429–444.

Javanmard, A. and Montanari, A. (2015). On online control of false discovery rate. *arXiv [stat.ME]*.

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

Pozzolo, A. D., Caelen, O., Johnson, R. A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166. ieeexplore.ieee.org.

Ramdas, A., Yang, F., Wainwright, M. J., and Jordan, M. I. (2017). Online control of the false discovery rate with decaying memory. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramdas, A., Zrnic, T., Wainwright, M., and Jordan, M. (2018). SAFFRON: an adaptive algorithm for online control of the false discovery rate. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4286–4294. PMLR.

Tian, J. and Ramdas, A. (2019). ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. *Adv. Neural Inf. Process. Syst.*, 32.

Xu, Z. and Ramdas, A. (2020). Dynamic algorithms for online multiple testing. *arXiv [stat.ME]*.

Zrnic, T., Jiang, D., Ramdas, A., and Jordan, M. (2020). The power of batching in multiple hypothesis testing. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3806–3815. PMLR.

Zrnic, T., Ramdas, A., and Jordan, M. I. (2021). Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.*, 22:33–31.

Appendix of Supplementary Materials

A Proofs

A.1 Proof of Monotonicity

In this section, we show that TOAD never reverses a previous rejection. This is formalized in the following lemma.

Lemma 1. *(No rejection reversals) The rejection sets produced by TOAD satisfy* $\mathcal{R}_t \subseteq \mathcal{R}_{t'}$ *for any* $0 < t < t' < \infty$.

Proof. We will show the result by induction. We start by showing that any hypothesis index j that is rejected at any stage t remains rejected at stage $t+1$ (i.e., if $j \in \mathcal{R}_t$ then $j \in \mathcal{R}_{t+1}$).

Let $\mathcal{A}_t = \mathcal{R}_t \cap \mathcal{C}_t \cap \mathcal{C}_{t+1}$ be the subset of indices rejected at time t that are active candidates at both stage t and $t+1$. Let $\mathcal{N}_t = \{\mathcal{R}_t \cap \mathcal{C}_t\} \backslash \mathcal{C}_{t+1}$ be the subset of indices rejected at time t that are active candidates at time t, but *not* at time $t+1$. Since $S_t = |\mathcal{C}_t \cap \mathcal{R}_t|$ is equal to the number of hypotheses that are both active and rejected at stage t, we have $S_t = |\mathcal{A}_t| + |\mathcal{N}_t|$.

Assume that $j \in \mathcal{R}_t$. If $j \notin \mathcal{C}_{t+1}$, then $j \in \mathcal{R}^{\text{old}}_{t+1}$, and the fact that $j \in \mathcal{R}_{t+1}$ holds automatically. Likewise if $j \notin \mathcal{C}_t$ then $j \notin \mathcal{C}_{t+1}$ (by definition) and the result again holds automatically. Otherwise, we must have $j \in \mathcal{A}_t$. So, hereafter, we assume that $j \in \mathcal{A}_t$.

We now consider two cases, depending on and whether or not adding W_{t+1} in stage $t+1$ changes the ordering of the test statistics $\{W_i\}_{i \in \mathcal{A}_t}$.

Case-1 $W_{t+1} \geq \max_{i \in \mathcal{A}_t} W_i$: Here, the test statistics $\{W_i\}_{i \in \mathcal{A}_t}$ advance to become the lowest $|\mathcal{A}_t|$ values in the new set of active statistics $\{W_i\}_{i \in \mathcal{C}_{t+1}}$, meaning that $W_{(|\mathcal{A}_t|,t+1)} = \max_{i \in \mathcal{A}_t} W_i$. Applying this, we have

$$\begin{aligned} W_{(|\mathcal{A}_t|,t+1)} = \max_{i \in \mathcal{A}_t} W_i \leq W_{(S_t,t)} \leq & \alpha\beta\left(S_t + |\mathcal{R}^{\text{old}}_t|\right) \\ = & \alpha\beta\left(|\mathcal{A}_t| + |\mathcal{N}_t| + |\mathcal{R}^{\text{old}}_t|\right) \\ = & \alpha\beta\left(|\mathcal{A}_t| + |\mathcal{R}^{\text{old}}_{t+1}|\right). \end{aligned} \tag{4}$$

Thus, from the definition of S_{T+1} (Eq (3)) we have $|\mathcal{A}_t| \leq S_{T+1}$. Finally, since

$$W_j \leq \max_{i \in \mathcal{A}_t} W_i = W_{(|\mathcal{A}_t|,t+1)} \leq W_{(S_{T+1},t+1)},$$

we see that j is rejected in stage $t+1$.

Case-2 $W_{t+1} < \max_{i \in \mathcal{A}_t} W_i$: Here, the test statistics $\{W_i\}_{i \in \mathcal{A}_t} \cup W_{t+1}$ become the lowest $(|\mathcal{A}_t|+1)$ values in the new set of active statistics $\{W_i\}_{i \in \mathcal{C}_{t+1}}$, and so $\max_{i \in \mathcal{A}_t} W_i = W_{(|\mathcal{A}_t|+1,t+1)}$. Following the same steps as above, we have

$$\begin{aligned} W_{(|\mathcal{A}_t|+1,t+1)} = \max_{i \in \mathcal{A}_t} W_i \leq W_{(S_t,t)} \leq & \alpha\beta\left(S_t + |\mathcal{R}^{\text{old}}_t|\right) \\ = & \alpha\beta\left(|\mathcal{A}_t| + |\mathcal{N}_t| + |\mathcal{R}^{\text{old}}_t|\right) \\ = & \alpha\beta\left(|\mathcal{A}_t| + |\mathcal{R}^{\text{old}}_{t+1}|\right) \\ \leq & \alpha\beta\left(|\mathcal{A}_t| + 1 + |\mathcal{R}^{\text{old}}_{t+1}|\right), \end{aligned}$$

where, in the last line, we use the fact that both shape functions and the identity function are nondecreasing. From the definition of S_{T+1} we have $|\mathcal{A}_t| + 1 \leq S_{T+1}$. Finally, since

$$W_j \leq \max_{i \in \mathcal{A}_t} W_i = W_{(|\mathcal{A}_t|+1,t+1)} \leq W_{(S_{T+1},t+1)},$$

we see that j is rejected in stage $t+1$.

Thus, for any $j \in \mathcal{R}_t$, we know that $j \in \mathcal{R}_{t+1}$. It now follows by induction that $j \in \mathcal{R}_{t'}$. □

A.2 Proof of Theorem 1

Part 1

Proof. Suppose the H_i is rejected at stage t, and let $T_i \leq t$ be the *first* stage at which H_i is rejected. We know that

$$W_i \leq W_{(S_{T_i}, T_i)} \leq \alpha\beta\left(S_{T_i} + |\mathcal{R}_{T_i}^{\text{old}}|\right) = \alpha\beta\left(|\mathcal{R}_{T_i}|\right) \leq \alpha\beta\left(1 \vee |\mathcal{R}_{T_i}|\right).$$

From Lemma 1, and from the fact that β (the identity function) is nondecreasing, we have

$$W_i \leq \alpha\beta\left(1 \vee |\mathcal{R}_{T_i}|\right) \leq \alpha\beta\left(1 \vee |\mathcal{R}_t|\right). \tag{5}$$

Finally, plugging in $W_i = P_i/A_i$, we see that rejecting any hypothesis H_i by time t requires

$$\begin{aligned} P_i/A_i &\leq \beta\left(1 \vee |\mathcal{R}_t|\right)\alpha \\ P_i &\leq \beta\left(1 \vee |\mathcal{R}_t|\right) A_i\alpha. \end{aligned}$$

We apply this fact in Line (6), below.

$$\begin{aligned} \text{FDR}(t) &= \mathbb{E}\left[\frac{\sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} 1\left(i \in \mathcal{R}_t\right)}{1 \vee |\mathcal{R}_t|}\right] \\ &\leq \mathbb{E}\left[\frac{\sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} 1\left(P_i \leq \beta(1 \vee |\mathcal{R}_t|)A_i\alpha\right)}{1 \vee |\mathcal{R}_t|}\right] \qquad (6) \\ &= \sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} \mathbb{E}\left[\frac{1\left(P_i \leq \beta(1 \vee |\mathcal{R}_t|)A_i\alpha\right)}{1 \vee |\mathcal{R}_t|}\right] \\ &= \sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} \mathbb{E}_{\mathcal{P}_{\tau_i}} \mathbb{E}\left[\frac{1\left(P_i \leq \beta(1 \vee |\mathcal{R}_t|)A_i\alpha\right)}{1 \vee |\mathcal{R}_t|} | \mathcal{P}_{\tau_i}\right]. \qquad (7) \end{aligned}$$

Within the inner expectation, Assumption 1 tells us that A_i is constant, and Assumption 2 tells us that P_i is stochastically lower bounded by a uniform variable on [0,1] (i.e., it is super-uniform conditional on $\mathcal{P}_{\tau_i}$). Using these facts along with Assumption 3, we can apply Lemma 3.2-ii from Blanchard and Roquain (2008) to see that the inner expectation is less than or equal to $A_i\alpha$.[1] Plugging this in and recalling that A_i is a function of $\mathcal{P}_{\tau_i}$, we have

$$\text{FDR}(t) \leq \sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} \mathbb{E}\left[A_i\alpha\right] = \alpha\mathbb{E}\left[\sum_{\{i \leq t\,:\, i \in \mathcal{H}_0\}} A_i\right] \leq \alpha,$$

where the last inequality follows from the definition of A_i.

As noted in Section G, below, all of these steps remain unchanged if we replace $\mathcal{P}_{\tau_i}$ throughout with $\mathcal{P}_{\tau_i}^{\text{obs}}$. □

Part 2

Proof. The proof is almost identical to the proof of Part 1. There are only two minor differences.

First, to show Eq (5), we must now cite the fact that shape functions are nondecreasing. The second difference comes where we previously used Assumption 3 to apply Lemma 3.2-ii from Blanchard and Roquain (2008) in order to show that the inner expectation in Eq (7) is less than or equal to $A_i\alpha$. Now, we instead use the fact that β is a shape function to apply Lemma 3.2-iii from Blanchard and Roquain (2008), which again shows that the inner expectation in Eq (7) is no more than $A_i\alpha$. □

[1] In applying this lemma, we set V, U, and c in Blanchard and Roquain's notation equal to $(1 \vee R_j)$, P_j, and $A_j\alpha$ in our notation, respectively.

B Power Comparison Between TOAD & $\text{Batch}_{\text{BH}}^{\text{PRDS}}$

We start by reviewing the $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ algorithm proposed in the appendix from Zrnic et al. (2020), and then move on to compare TOAD against $\text{Batch}_{\text{BH}}^{\text{PRDS}}$. As the name suggests, $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ starts by breaking down the sequence of p-values into "batches." For each batch b, the authors apply BH to the b^{th} batch at an alpha level of $\alpha^{(b)}$. Let $n^{(b)}$ denote the size of the b^{th} batch, and let $R^{(b)}$ be the number of hypotheses rejected from batch b. Here, we use superscript notation to help distinguish between batch indices and test indices. Zrnic et al. define a $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ procedure as any method of defining alpha levels $\{\alpha^{(b)}\}_{b=1}^{\infty}$ that satisfies

$$\sum_{s \leq b} \alpha^{(s)} \frac{n^{(s)}}{n^{(s)} + \sum_{r<s} R^{(r)}} \leq \alpha \tag{8}$$

for all batches b, where α is the desired FDR control level. Additionally, Zrnic et al. require each $\alpha^{(s)}$ to depend only on the p-values from the preceding batches.

In order to compare TOAD and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$, we first need to translate the above "batch" notation into the more general notation of "deadlines." Given a sequence of p-values $P_1, P_2, \ldots$, let $g_1, g_2, \ldots$ be "group" or "batch" labels for each p-value. For example, if we observe two batches, each of size two, then $(g_1, g_2, g_3, g_4) = (1, 1, 2, 2)$. Again, we generally use subscript "stage indices" to refer to the indices of hypotheses and tests, and use superscript "batch indices" to denote batch numbers. We define each deadline parameter $d_i = \max\{i' : g_i = g_{i'}\}$ to be the (stage) index of the last test that is in the same batch as P_i.

If batch b ends at stage t, then the following notational equivalencies can be made.

- $g_t = b$ is the batch label for stage t (i.e., for the t^{th} hypothesis);
- $\mathcal{C}_t = \{i \leq t : g_i = g_t\}$ is the set of stage indices in the b^{th} batch; and
- $n^{(b)} = n^{(g_t)} = |\mathcal{C}_t|$ is the size of batch b.

Equipped with this notation, we can formally compare TOAD and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$.

Remark 1. *(Power comparison) Given a sequence of alpha levels $\{\alpha^{(b)}\}_{b=1}^{\infty}$ used by the Batch_{BH}^{PRDS} method, if we set β to be the identity function and set*

$$A_i = \frac{\alpha^{(g_i)}}{\alpha\left(n^{(g_i)} + \sum_{s<g_i} R^{(s)}\right)},$$

for all i, then any hypothesis rejected by Batch_{BH}^{PRDS} is also rejected by TOAD.

Proof of Remark 1

Proof. First, we note that this choice of A_i still satisfies our requirement that A_i is a function of the preceding p-values. We also note that, for any stage t at which a batch ends, we have

$$\begin{aligned}
\sum_{i=1}^{t} A_i &= \sum_{i=1}^{t} \frac{\alpha^{(g_i)}}{\alpha\left(n^{(g_i)} + \sum_{s<g_i} R^{(s)}\right)} \\
&= \sum_{b=1}^{g_t} n^{(b)} \frac{\alpha^{(b)}}{\alpha\left(n^{(b)} + \sum_{s<b} R^{(s)}\right)} \\
&= \frac{1}{\alpha} \sum_{b=1}^{g_t} \alpha^{(b)} \frac{n^{(b)}}{\left(n^{(b)} + \sum_{s<b} R^{(s)}\right)} \\
&\leq 1,
\end{aligned}$$

where the last line comes from Eq (8).

Next, we write a more explicit version of the $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ procedure that is closer in format to Algorithm 1.

Algorithm 2. (Alternative description Batch_{BH}^{PRDS}) Take as input an alpha level for the first batch, denoted by $\alpha^{(1)}$.

For each stage t:

1. *Let $\mathcal{R}_t^{old}$ be the set of hypotheses rejected in previous batches.*
2. *If the current batch does not end at stage t (i.e., if $t < d_t$): reject $\mathcal{R}_t = \mathcal{R}_t^{old}$.*
3. *If the current batch ends at stage t (i.e., if $t = d_t$):*
   (a) *Recall that $\mathcal{C}_t = \{i \le t : g_i = g_t\}$ is the set of hypothesis indices in the same batch as H_t, so that $n^{(g_t)} = |\mathcal{C}_t|$.*
   (b) *Let $P_{(j,t)}$ denote the j^{th} highest p-value from the set $\{P_i\}_{i\in\mathcal{C}_t}$, that is, the j^{th} highest p-value in the current batch.*
   (c) *Set $\tilde{S}_t = \max\left\{j \le |\mathcal{C}_t| : P_{(j,t)} \le \frac{j}{|\mathcal{C}_t|}\alpha^{(g_t)}\right\}$.*
   (d) *Reject $\mathcal{R}_t = \mathcal{R}_t^{old} \cup \{i \in \mathcal{C}_t : P_i \le P_{(\tilde{S}_t,t)}\}$.*
   (e) *Define the alpha level $\alpha^{(t+1)}$ to be used in the next batch, in accordance with Eq (8).*

Above, all we have done is changed the notation to define indices at the test level instead of the batch level, and plugged in the steps of the BH procedure.

For all tests i, let $W_i = P_i/A_i$. Let $W_{(j,t)}$ to be the j^{th} highest value in the set $\{W_i\}_{i\in\mathcal{C}_t}$. Since A_i is constant within a batch $\mathcal{C}_t$, we know that W_i is proportional to P_i among the indices $i \in \mathcal{C}_t$. Thus, $W_{(j,t)} = P_{(j,t)}/A_t$. This leads to an equivalent definition of $\tilde{S}_t$:

$$\begin{aligned}
\tilde{S}_t &= \max\left\{j \le |\mathcal{C}_t| : P_{(j,t)} \le \frac{j}{|\mathcal{C}_t|}\alpha^{(g_t)}\right\} \\
&= \max\left\{j \le |\mathcal{C}_t| : P_{(j,t)}/A_t \le \frac{j\alpha^{(g_t)}}{|\mathcal{C}_t|}A_t^{-1}\right\} \\
&= \max\left\{j \le |\mathcal{C}_t| : W_{(j,t)} \le \frac{j\alpha^{(g_t)}}{|\mathcal{C}_t|} \times \frac{\alpha\left(n^{(g_t)} + \sum_{s<g_t} R^{(s)}\right)}{\alpha^{(g_t)}}\right\} \\
&= \max\left\{j \le |\mathcal{C}_t| : W_{(j,t)} \le \frac{j\alpha}{|\mathcal{C}_t|}\left(n^{(g_t)} + \sum_{s<g_t} R^{(s)}\right)\right\} \\
&= \max\left\{j \le |\mathcal{C}_t| : W_{(j,t)} \le \frac{j\alpha}{|\mathcal{C}_t|}\left(|\mathcal{C}_t| + |\mathcal{R}_t^{\text{old}}|\right)\right\} \\
&= \max\left\{j \le |\mathcal{C}_t| : W_{(j,t)} \le \alpha\left(j + \frac{j}{|\mathcal{C}_t|} \times |\mathcal{R}_t^{\text{old}}|\right)\right\}.
\end{aligned}$$

Thus, Step 3d is equivalent to rejecting $\mathcal{R}_t^{\text{old}} \cup \{i \in \mathcal{C}_t : W_i \le W_{(\tilde{S}_t,t)}\}$. All that remains is to show that $\tilde{S}_t \le S_t$, where S_t is the value in Eq (3). This follows immediately from the fact that $j/|\mathcal{C}_t| \le 1$. □

C Sufficient Conditions for Assumption 3

First, we review the definition of positive regression dependence on a subset (PRDS; Benjamini and Yekutieli, 2001). Given a value $t \in \mathbb{N}$, let $\mathbf{P} = (P_1, \ldots, P_t)$ (we suppress the dependence on t for brevity). We say that a set of k-dimensional vectors $D \in [0,1]^k$ is *increasing* if, for any vector $\mathbf{x} = (x_1, \ldots, x_t) \in D$ and any vector $\mathbf{y} = (y_1, \ldots, y_t)$ satisfying $x_i \le y_i$ for all i, it must also be that $\mathbf{y} \in D$. Given a subset $I_0 \subseteq \{1, \ldots t\}$, the PRDS condition states that, for any increasing set D and any $j \in I_0$, the conditional probability $\mathbb{P}(\mathbf{P} \in D | P_j = u)$ is nondecreasing in u.

Following the argument of Blanchard and Roquain (2008; see Proposition 3.6 and citations therein), we outline two sufficient conditions that imply Assumption 3.

Condition 1. *For any $t \in \mathbb{N}$, increasing any of the first t p-values cannot increase the total number of discoveries produced by time t.*

Condition 2. *For any $j, t \in \mathbb{N}$ such that $j \leq t$ and $H_j \in \mathcal{H}_0$, and for any increasing set $D \in [0,1]^t$, the conditional probability $\mathbb{P}(\mathbf{P} \in D | P_j = u, \mathcal{P}_{\tau_j})$ is nondecreasing in u.*

Condition 1 holds, for example, if all parameters A_i are prespecified, or if they are monotonically decreasing in the p-values observed so far. Alternatively, Condition 1 holds if we set $A_1, A_2, \ldots$ equal to a predetermined sequence of constants $a_1, a_2, \ldots$ until a certain number of discoveries are found, and set $A_t = 0$ afterwards. Condition 2 is a modified version of the PRDS requirement.

Remark 2. *If Conditions 1 & 2 hold then Assumption 3 holds.*

Proof of Remark 2

Proof. (Adapted from Blanchard and Roquain, 2008) For any $0 \leq r \leq t$, let $D_r \subseteq [0,1]^t$ be the set of all possible p-values that produce no more than $1 \vee r$ discoveries by stage t. Since increasing any p-value will not increase the number of discoveries by stage t, we know that D_r is an increasing set. For any $u < u'$, let $\gamma = P(P_j \leq u | P_j \leq u', \mathcal{P}_{\tau_j})$. Then

$$
\begin{aligned}
&\mathbb{P}(1 \vee |\mathcal{R}_t| \leq r | P_j \leq u', \mathcal{P}_{\tau_j}) \\
&= \mathbb{P}(\mathbf{P} \in D_r | P_j \leq u', \mathcal{P}_{\tau_j}) && (9) \\
&= \mathbb{E}\left[\mathbb{P}(\mathbf{P} \in D_r | P_j, \mathcal{P}_{\tau_j}) \quad | \quad P_j \leq u', \mathcal{P}_{\tau_j}\right] \\
&= \quad \mathbb{E}\left[\mathbb{P}(\mathbf{P} \in D_r | P_j, \mathcal{P}_{\tau_j}) \quad | \quad P_j \leq u, \mathcal{P}_{\tau_j}\right] \gamma && (10) \\
&\quad + \mathbb{E}\left[\mathbb{P}(\mathbf{P} \in D_r | P_j, \mathcal{P}_{\tau_j}) \quad | \quad u < P_j \leq u', \mathcal{P}_{\tau_j}\right] (1-\gamma). && (11)
\end{aligned}
$$

Under Condition 2, the expectation in Line (10) is smaller than the expectation in Line (11). Thus, if we were to replace the expectation in Line (11) with the expectation in Line (10), the sum shown in Lines (10)-(11) would be reduced. Making this substitution and combining terms gives

$$
\begin{aligned}
\mathbb{P}(1 \vee |\mathcal{R}_t| \leq r | P_j \leq u', \mathcal{P}_{\tau_j}) &\geq \mathbb{E}\left[\mathbb{P}(\mathbf{P} \in D_r | P_j, \mathcal{P}_{\tau_j}) \quad | \quad P_j \leq u, \mathcal{P}_{\tau_j}\right] \\
&= \mathbb{P}(\mathbf{P} \in D_r | P_j \leq u, \mathcal{P}_{\tau_j}) \\
&= \mathbb{P}(1 \vee |\mathcal{R}_t| \leq r | P_j \leq u, \mathcal{P}_{\tau_j}).
\end{aligned}
$$

This proves the result. □

D FDR for Naive-BH

Here, we show that Naive-BH controls the FDR whenever the p-values are PRDS (see Section C of this appendix; and Benjamini and Yekutieli, 2001) on the subset of test statistics corresponding to the true null hypotheses.

Let $R^{(b)}$ be the number of rejections from the b^{th} batch and let $V^{(b)}$ denote the number of erroneous rejections from the b^{th} batch. Note that $\lfloor t/n_{\text{batch}} \rfloor$ is the number of batches that have completed by stage t. Recall also that Naive-BH uses an alpha level of $\alpha(t_{\max}/n_{\text{batch}})^{-1}$ for each batch. We have

$$
\begin{aligned}
FDR(t) = \mathbb{E}\left[\frac{\sum_{\{b \leq \lfloor t/n_{\text{batch}} \rfloor\}} V^{(b)}}{1 \vee \sum_{\{b' \leq \lfloor t/n_{\text{batch}} \rfloor\}} R^{(b')}}\right] &= \sum_{\{b \leq \lfloor t/n_{\text{batch}} \rfloor\}} \mathbb{E}\left[\frac{V^{(b)}}{1 \vee \sum_{\{b' \leq \lfloor t/n_{\text{batch}} \rfloor\}} R^{(b')}}\right] \\
&\leq \sum_{\{b \leq \lfloor t/n_{\text{batch}} \rfloor\}} \mathbb{E}\left[\frac{V^{(b)}}{1 \vee R^{(b)}}\right] \\
&\leq \sum_{\{b \leq \lfloor t/n_{\text{batch}} \rfloor\}} \alpha \frac{n_{\text{batch}}}{t_{\max}} && (12) \\
&\leq \frac{t_{\max}}{n_{\text{batch}}} \times \alpha \frac{n_{\text{batch}}}{t_{\max}} \\
&= \alpha,
\end{aligned}
$$

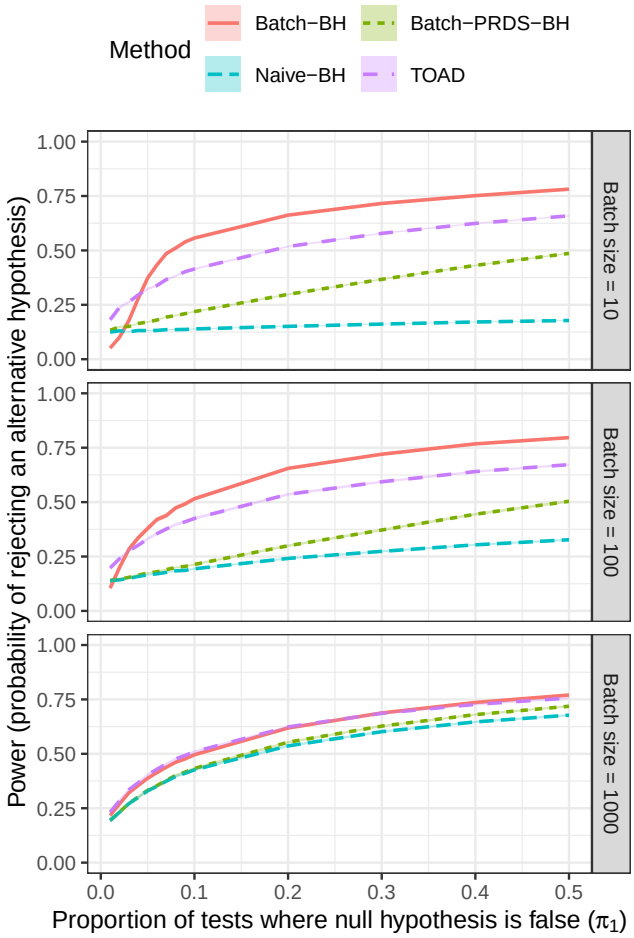

Figure 3: Simulated power under AR1 correlation structure

where Line (12) comes from the well-known result that BH controls the (within batch) FDR under the PRDS assumption (Theorem 1.2 from Benjamini and Yekutieli, 2001).

E Additional Simulations

To expand our simulations, we first consider the setting where test statistics follow a first order autoregressive (AR1) structure within each batch. Namely, for any two indices (i, i') within the same batch, we simulate the test statistics $Z_i, Z_{i'}$ such that $Var(Z_i) = Var(Z_{i'}) = 1$, and $Cov(Z_i, Z_{i'}) = 0.9^{\frac{|i-i'|}{2}}$.

Next, we consider the setting described by Zrnic et al. (2020) and Javanmard and Montanari (2018) in which the amount of signal can vary drastically across test statistics. For each index i associated with an alternative distribution, we draw μ_i from a random normal distribution with mean zero and variance $2\log(t_{\max}) \approx 16$. Here, we set $P_i = 2\Phi(-|Z_i|)$ to be the p-value resulting from a two-sided test of H_i.

Figures 3, 4, 5 & 6 show the results of these simulations. Overall, the pattern resembles what we observe in our main simulations. However, for the case of random mean parameters, both the differences in power and the degree of FDR inflation are less pronounced. For the AR1 case, the FDR inflation of Batch_{BH} is more pronounced.

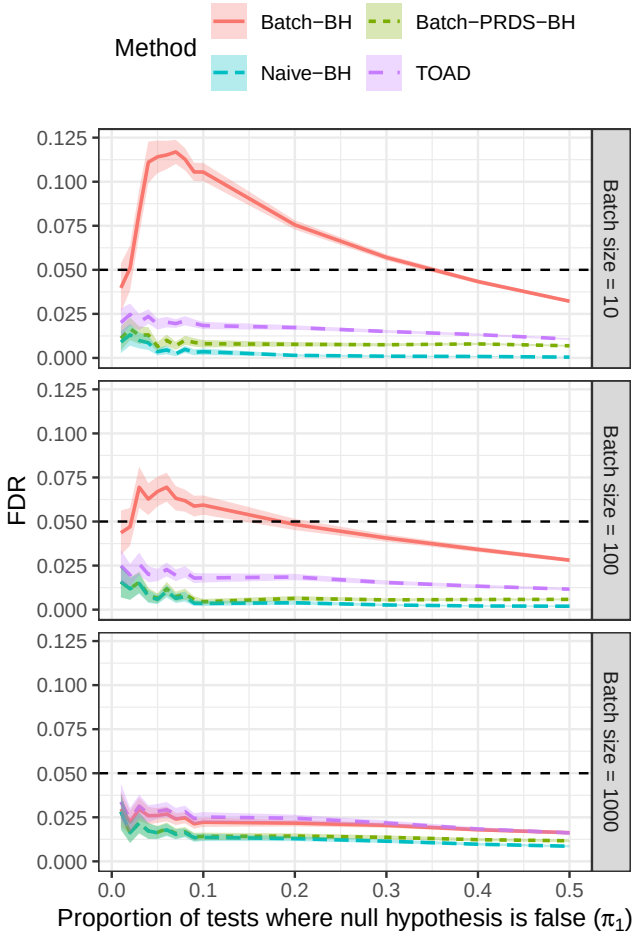

Figure 4: Simulated FDR under AR1 correlation structure

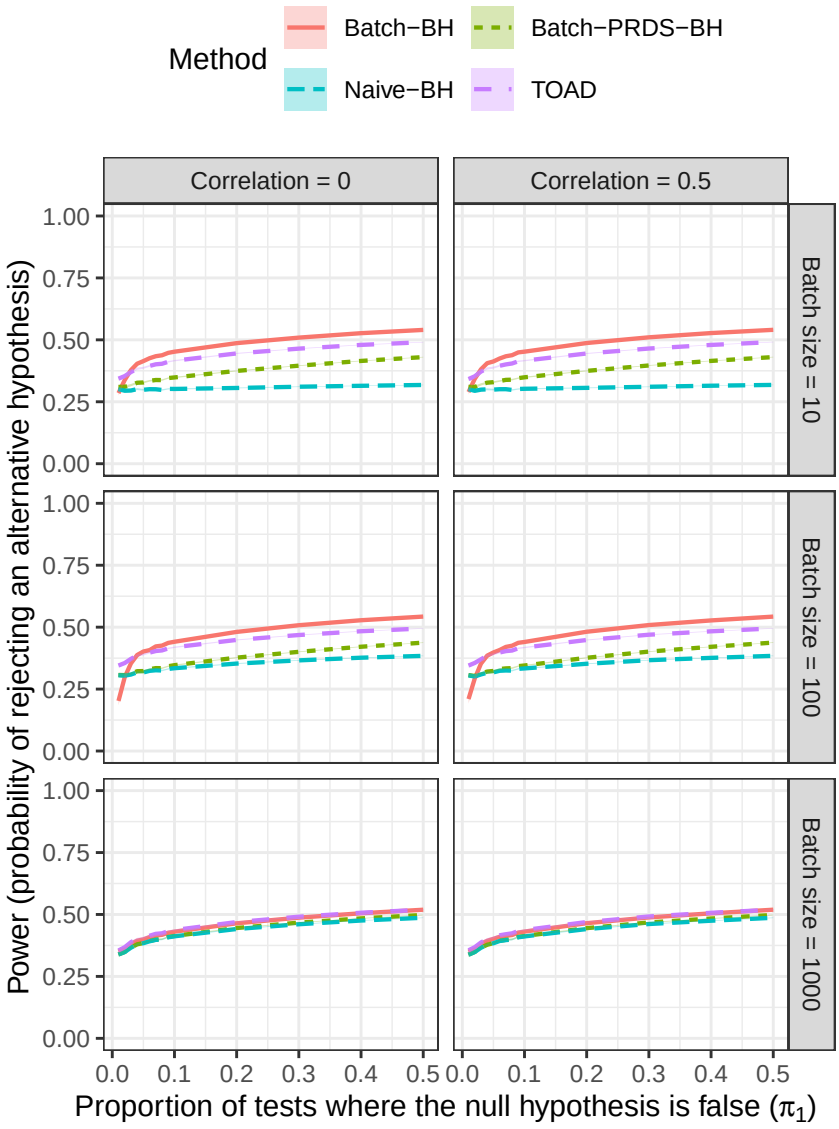

Figure 5: Simulated power under randomly selected values for the mean parameters

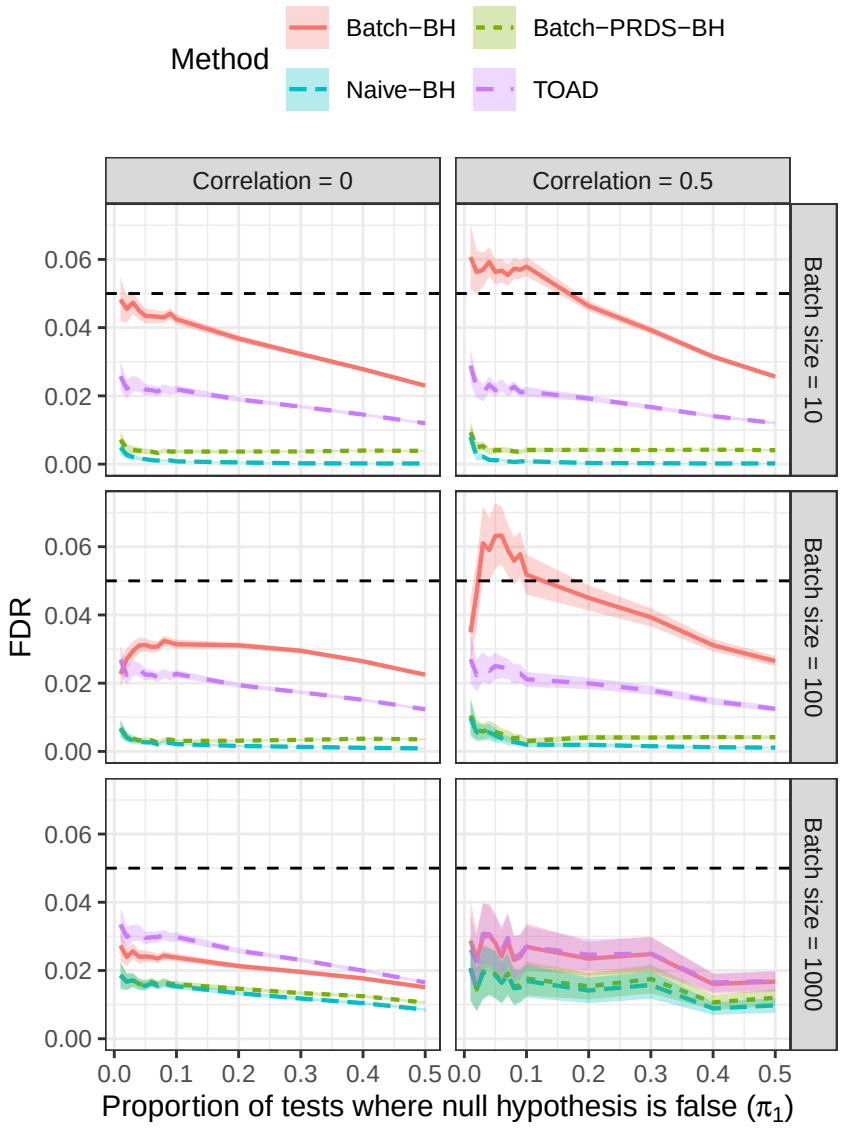

Figure 6: Simulated FDR under randomly selected values for the mean parameters

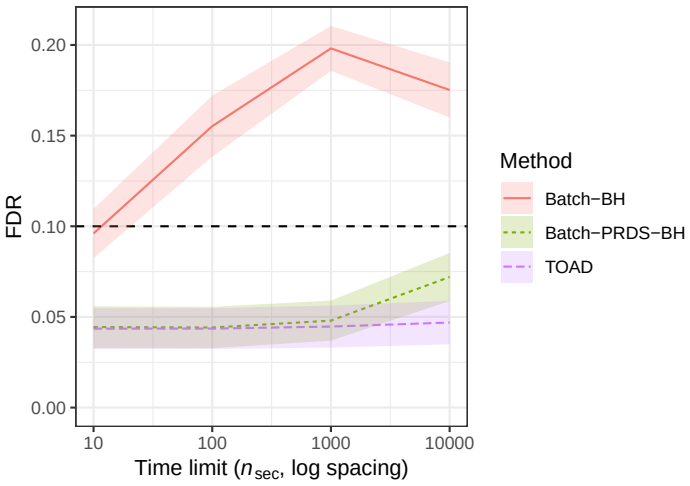

Figure 7: FDR for credit card fraud dataset

F Example Analysis of Fraud Detection in Credit Card Transaction Data

Next, we study a dataset containing 48 hours of credit card transactions (Pozzolo et al., 2015),[2] which was also studied by Zrnic et al. (2020). Each transaction is time-stamped and labeled as either fraudulent or not. Rather that including features such as price and location for each transaction, the dataset instead contains the first 28 principal components generated from such features, in order to preserve confidentiality.

In this setting, each transaction i is associated with a null hypothesis H_i that the transaction is credible. Rejecting this hypothesis amounts to flagging the transaction as fraudulent.

To generate p-values for each hypothesis, we follow the same procedure as Zrnic et al., 2020. We first randomly partition the dataset into three subsets. Using the first subset (60% of transactions), we train a logistic regression model applied to all principle components and the time-stamp, with fraud status as the outcome. The output of this model, the predicted probability of a transaction being fraudulent, will serve as our test statistic for each hypothesis. Next, using the second subset (20% of transactions), we apply this model to all non-fraudulent transactions to obtain a null distribution for our test statistic. Finally, for each of the remaining transactions, we use our model to generate a test statistic, the estimated the probability of fraud, and compare this test statistic against our null distribution to generate a p-value.

Given a series of $t_{\max}$ such p-values, we apply TOAD with constant tuning parameters $A_1 = A_2 = \cdots = A_{t_{\max}} = 1/t_{\max}$ and an alpha level of 0.1. Let T_i be the time-stamp for the i^{th} transaction (in seconds). We define n_{sec} to be a tuning parameter equal to the number seconds we have to make a decision about each transaction, and define each decision deadline as $d_i = \max\{i' : |T_{i'} - T_i| \leq n_{\text{sec}}\}$. That is, d_i is the index of the last hypothesis that is tested within n_{sec} seconds of H_i. We consider tuning parameter values $n_{\text{sec}} \in \{10, 100, 1000, 1000\}$. In order to estimate FDR and power for each setting, we repeat the procedure over 100 different, random training partitions and average the results.

As comparator methods, we again consider the Batch_{BH} and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ algorithms (Zrnic et al., 2020). We define the "group" or "batch" label for H_1 as $g_1 = 1$, and define subsequent batch labels sequentially, as $g_i = g_{i-1}$ if $T_i - \min_{\{i' \leq i : g_{i'} = g_i\}} T_{i'} \leq n_{\text{sec}}$ and $g_i = g_{i-1} + 1$ otherwise. Thus, each batch lasts for no more than n_{sec} seconds, although there may be time gaps between batches. This ensures that, for any transaction, Batch_{BH} and $\text{Batch}_{\text{BH}}^{\text{PRDS}}$ both deliver a final decision within n_{sec} seconds of that transaction.

Figures 7 and 8 show the results, which are comparable to those of our simulations. Batch_{BH} typically produces the highest power, but can inflate FDR by up to a factor of 2. Of the methods that control FDR at the desired rate, TOAD produces the highest power.

[2] https://www.kaggle.com/mlg-ulb/creditcardfraud

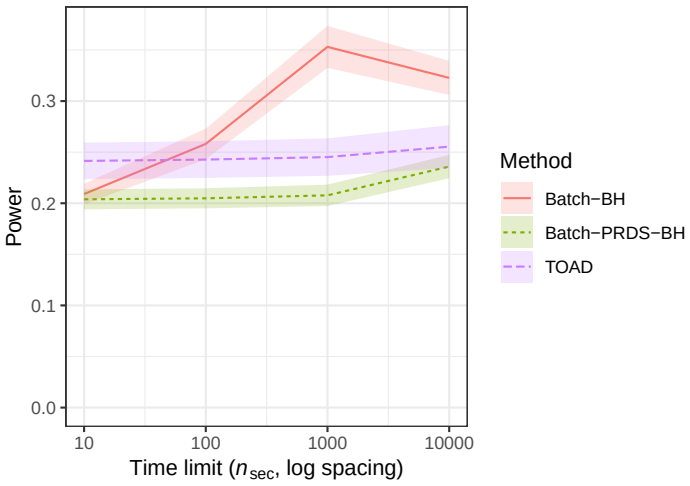

Figure 8: Power for credit card fraud dataset

Table 1: Online Hypothesis Reordering

STAGE (t)	H_t	**OPTION 1 FOR** A_t	**OPTION 2 FOR** A_t
1	$\tilde{H}^{(1)}$	1/3	1/3
2	$\tilde{H}^{(2)}$	1/3	0
3	$\tilde{H}^{(3)}$	0	1/3
4	$\tilde{H}^{(2)}$	0	1/3
5	$\tilde{H}^{(3)}$	1/3	0

Table 1 Caption: The first column shows the stage index for a 5-stage experiment. The second column shows a sequence of hypotheses, including duplicates, to be tested in an online fashion at each stage. The third and fourth columns offer different choices for the tuning parameters $A_2, \ldots, A_5$, where the choice between these options can be made at the end of Stage 1 (i.e., after observing P_1). Option 1 amounts to testing the hypotheses in the order $\tilde{H}^{(1)}, \tilde{H}^{(2)}, \tilde{H}^{(3)}$, while Option 2 amounts to testing the hypotheses in the order $\tilde{H}^{(1)}, \tilde{H}^{(3)}, \tilde{H}^{(2)}$.

G Ignoring Hypotheses, and Adaptive Hypothesis Reordering

A central advantage of online procedures is their ability to selectively ignore hypotheses based on preliminary results. Here, we say that a hypothesis H_i is "ignored" if $A_i = 0$ (see also Appendix B of Ramdas et al., 2017 for a similar discussion). Using the idea of ignoring hypotheses as a building block, we can quickly encompass other types of online strategies. For example, if the hypothesis sequence $H_1, H_2, \ldots$ is sufficiently diverse, then we can effectively *define* our hypotheses adaptively by ignoring those hypotheses that are no longer of interest.

Similarly, ignoring hypotheses effectively lets us adaptively *reorder* the available hypotheses. For example, suppose that a researcher plans to test three unique hypotheses $\tilde{H}^{(1)}, \tilde{H}^{(2)}, \tilde{H}^{(3)}$, but wishes to test the last two in an adaptive order. This can be achieved by defining the expanded, 5-stage hypothesis sequence

$$(H_1, H_2, H_3, H_4, H_5) = (\tilde{H}^{(1)}, \tilde{H}^{(2)}, \tilde{H}^{(3)}, \tilde{H}^{(2)}, \tilde{H}^{(3)}),$$

shown in Table 1. From here, depending on how the parameters (A_2, A_3, A_4, A_5) are selected, the researcher can use the result of the first test to decide whether to test $\tilde{H}^{(2)}$ before $\tilde{H}^{(3)}$, or vice versa (see details in Table 1). The same approach can be used to reorder arbitrarily large hypothesis sets.

In order to leverage the benefits of ignoring hypotheses, we will need restrict the information used to define upcoming threshold parameters A_i. At present, our Assumption 2 requires that future test statistics be conditionally uniform given the previous p-values, and such a condition can be impossible to satisfy if the hypothesis sequence

contains repeats. For this reason, we suggest modifying Assumptions 1, 2 & 3 so that testing decisions depend only on the previous "unignored" hypotheses. To formalize this, we define $P_t^{\text{obs}} = P_t \times 1(A_t > 0) - 1(A_t = 0)$ to be equal to -1 if H_t is ignored and equal to P_t otherwise. Thus, the sequence $\mathcal{P}_{\tau_i}^{\text{obs}} = \{P_{i'}^{\text{obs}}\}_{i' \leq \tau_i}$ contains the information in the first τ_i p-values that is not ignored. Our Theorem 1 is unchanged if we replace $\mathcal{P}_{\tau_i}$ with $\mathcal{P}_{\tau_i}^{\text{obs}}$ in Assumptions 1, 2 & 3 (see the proof of Theorem 1).

H Forgetting Antiquated Tests

Proof. Here, we show that Theorem 1 still holds if we replace $\text{FDR}(t)$ with $\text{FDR}_{\text{recent}}(t)$; replace $\mathcal{R}_t^{\text{old}}$ with the empty set $\emptyset$ throughout the procedure; and relax the requirement that $\sum_{i=1}^{\infty} A_i \leq 1$ to instead require that $\sum_{i \in \mathcal{C}_t} A_i \leq 1$ for all t.

To show Part 1, suppose the H_i is rejected at stage t. We know that

$$P_i/A_i = W_i \leq W_{(S_t,t)} \leq \alpha\beta\left(S_t\right) = \alpha\beta\left(|\mathcal{R}_t|\right) \leq \alpha\beta\left(1 \vee |\mathcal{R}_t|\right).$$

Thus, rejecting any hypothesis H_i at time t requires that

$$P_i \leq \beta\left(1 \vee |\mathcal{R}_t|\right) A_i \alpha.$$

We applying this fact, we have

$$\begin{aligned}\text{FDR}_{\text{recent}}(t) &= \mathbb{E}\left[\frac{\sum_{\{i \leq t\,:\, i \in \mathcal{H}_0 \cap \mathcal{C}_t\}} 1\left(i \in \mathcal{R}_t\right)}{1 \vee |\mathcal{R}_t \cap \mathcal{C}_t|}\right] \\ &\leq \sum_{\{i \leq t\,:\, i \in \mathcal{H}_0 \cap \mathcal{C}_t\}} \mathbb{E}_{\mathcal{P}_{\tau_i}} \mathbb{E}\left[\frac{1\left(P_i \leq \beta(1 \vee R_t) A_i \alpha\right)}{1 \vee |\mathcal{R}_t \cap \mathcal{C}_t|} | \mathcal{P}_{\tau_i}\right]. \end{aligned} \tag{13}$$

As in Appendix A.2, we apply Lemma 3.2-ii from Blanchard and Roquain (2008) to see that the inner expectation is no more than $A_i\alpha$. Thus,

$$\text{FDR}_{\text{recent}}(t) \leq \sum_{\{i \leq t\,:\, i \in \mathcal{H}_0 \cap \mathcal{C}_t\}} \mathbb{E}\left[A_i \alpha\right] = \alpha \mathbb{E}\left[\sum_{\{i \leq t\,:\, i \in \mathcal{H}_0 \cap \mathcal{C}_t\}} A_i\right] \leq \alpha.$$

To show Part 2, we follow the same steps with the exception of applying Lemma 3.2-iii from Blanchard and Roquain (2008) to Eq (13). □