SAFFRON: an Adaptive Algorithm for
Online Control of the False Discovery Rate

Aaditya Ramdas [1]   Tijana Zrnic [2]   Martin J. Wainwright [1]   Michael I. Jordan [1]

Abstract

In the online false discovery rate (FDR) problem, one observes a possibly infinite sequence of p-values $P_1, P_2, \ldots$, each testing a different null hypothesis, and an algorithm must pick a sequence of rejection thresholds $\alpha_1, \alpha_2, \ldots$ in an online fashion, effectively rejecting the k-th null hypothesis whenever $P_k \leq \alpha_k$. Importantly, α_k must be a function of the past, and cannot depend on P_k or any of the later unseen p-values, and must be chosen to guarantee that for any time t, the FDR up to time t is less than some pre-determined quantity $\alpha \in (0, 1)$. In this work, we present a powerful new framework for online FDR control that we refer to as "SAFFRON". Like older alpha-investing algorithms, SAFFRON starts off with an error budget (called alpha-wealth) that it intelligently allocates to different tests over time, earning back some alpha-wealth whenever it makes a new discovery. However, unlike older methods, SAFFRON's threshold sequence is based on a novel estimate of the alpha fraction that it allocates to true null hypotheses. In the offline setting, algorithms that employ an estimate of the proportion of true nulls are called "adaptive", hence SAFFRON can be seen as an online analogue of the offline Storey-BH adaptive procedure. Just as Storey-BH is typically more powerful than the Benjamini-Hochberg (BH) procedure under independence, we demonstrate that SAFFRON is also more powerful than its non-adaptive counterparts such as LORD.

[1]Departments of Statistics and Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA. Correspondence to: Aaditya Ramdas <aramdas@eecs.berkeley.edu>, Tijana Zrnic <tijana@eecs.berkeley.edu>, Martin J. Wainwright <wainwrig@eecs.berkeley.edu>, Michael I. Jordan <jordan@eecs.berkeley.edu>.

1. Introduction

It is now commonplace in science and technology to make thousands or even millions of related decisions based on data analysis. As a simplified example, to discover which genes may be related to diabetes, we can formulate the decision-making problem in terms of hypotheses that take the form "gene X is not associated with diabetes," for many different genes X, and test for which of these null hypotheses can be confidently rejected by the data. As first identified by Tukey in a seminal 1953 manuscript (1953), the central difficulty when testing a large number of null hypotheses is that several of them may appear to be false, purely by chance. Arguably, we would like the set of rejected null hypotheses $\mathcal{R}$ to have high *precision*, so that most discovered genes are indeed truly correlated with diabetes and further investigations are not fruitless. Unfortunately, separately controlling the false positive rate for each individual test actually does not provide any guarantee on the precision. This motivated the development of procedures that can provide guarantees on an error metric called the false discovery rate (FDR) (Benjamini & Hochberg, 1995), defined as:

$$\text{FDR} \equiv \mathbb{E}\left[\text{FDP}(\mathcal{R})\right] = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}|}{|\mathcal{R}|}\right],$$

where $\mathcal{H}^0$ is the unknown set of truly null hypotheses, and $0/0 \equiv 0$. Here the FDP represents the ratio of falsely rejected nulls to the total number of rejected nulls, and since the set of discoveries $\mathcal{R}$ is data-dependent, the FDR takes an expectation over the underlying randomness. The evidence from a hypothesis test can typically be summarized in terms of a p-value, and so offline multiple testing algorithms take a set of p-values $\{P_i\}$ as their input, and a target FDR level $\alpha \in (0, 1)$, and produce a rejected set $\mathcal{R}$ that is guaranteed to have FDR $\leq \alpha$. Of course, one also desires a high recall, or equivalently a low false negative rate, but without assumptions on many uncontrollable factors like the frequency and strength of signals, additional guarantees on the recall are impossible.

While the offline paradigm previously described is the classical setting for multiple decision-making, the corresponding online problem is emerging as a major area of its own. For example, large information technology companies run thou-

sands of A/B tests every week of the year, and decisions about whether or not to reject the corresponding null hypothesis must be made without knowing the outcomes of future tests; indeed, future null hypotheses may depend on the outcome of the current test. The current standard of setting all thresholds α_k to a fixed quantity such as 0.05 does not provide any control of the FDR. Hence, the following hypothetical scenario is entirely plausible: a company conducts 1000 tests in one week, each with a target false positive rate of 0.05; it happens to make 80 discoveries in total of which 50 are accidental false discoveries, ending up with an FDP of $5/8$. Such uncontrolled error rates can have severe financial and social consequences.

The first method for online control of the FDR was the alpha-investing algorithm of Foster and Stine (2008), later extended to generalized alpha-investing (GAI) algorithms by Aharoni and Rosset (2014). Recently, Javanmard and Montanari (2017) proposed variants of GAI algorithms that control the FDR (as opposed to the modified FDR controlled in the original paper (Foster & Stine, 2008)), including a new algorithm called LORD. The GAI++ algorithms by Ramdas et al. (2017) improved the earlier GAI algorithms (uniformly), and the improved LORD++ (henceforth LORD) method arguably represents the current state-of-the-art in online multiple hypothesis testing.

The current paper's central contribution is the derivation and analysis of a powerful new class of online FDR algorithms called "SAFFRON" (Serial estimate of the Alpha Fraction that is Futilely Rationed On true Null hypotheses). As an instance of the GAI framework, the SAFFRON method starts off with an error budget, referred to as *alpha-wealth*, that it allocates to different tests over time, earning back some alpha-wealth whenever it makes a new discovery. However, unlike earlier work in the online setting, SAFFRON is an adaptive method, meaning that it is based on an estimate of the proportion of true nulls. In the offline setting, adaptive methods were proposed by Storey (2002; 2004), who showed that they are more powerful than the Benjamini-Hochberg (BH) procedure (1995) under independence assumptions; this advantage usually increases with the proportion of non-nulls and the signal strength. Thus, the SAFFRON method can be viewed as an online analogue of Storey's adaptive version of the BH procedure. As shown in Figure 1, our simulations show that SAFFRON demonstrates the same types of advantages over its non-adaptive counterparts, such as LORD and alpha-investing. Furthermore, the ideas behind SAFFRON's derivation can provide a natural template for the design and analysis of a suite of other adaptive online methods.

The rest of this paper is organized as follows. In Section 2, we derive the SAFFRON algorithm from first principles, leaving the precise statement and the proof of a central tech-
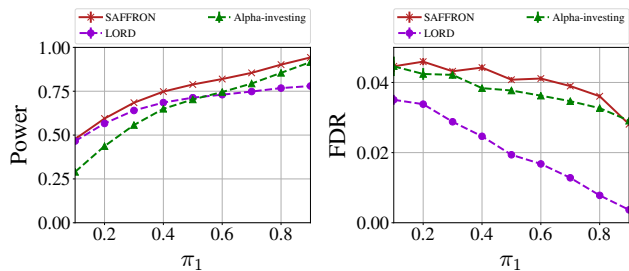

Figure 1. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON, LORD and alpha-investing at target FDR level $\alpha = 0.05$. The p-values are drawn as $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian CDF, and $Z_i \sim N(\mu_i, 1)$, where nulls have $\mu_i = 0$ and non-nulls have $\mu_i \sim N(3, 1)$.

nical lemma for Section 3. In Section 4, we investigate the practical choice of tuning parameters, and demonstrate the effectiveness of our recommended choice using simulations. We end with a summary in Section 5.

2. Deriving the SAFFRON Algorithm

Before deriving the SAFFRON algorithm, it is useful to recap a few concepts. By definition of a p-value, if the hypothesis H_i is truly null, then the corresponding p-value is stochastically larger than the uniform distribution ("super-uniformly distributed," for short), meaning that:

$$\text{If the null hypothesis } H_i \text{ is true, then} \quad \Pr\{P_i \leq u\} \leq u \text{ for all } u \in [0,1]. \tag{1}$$

For any online FDR procedure, let the rejected set after t steps be denoted by $\mathcal{R}(t)$. More precisely, this set consists of all p-values among the first t ones for which the indicator for rejection is equal to 1; i.e., $R_j := \mathbf{1}\{P_j \leq \alpha_j\} = 1$, for all $j \leq t$. While we have already defined the classical FDP and FDR in the introduction, several authors, including Foster and Stine (2008), have considered a modified FDR, defined as:

$$\text{mFDR}(t) := \frac{\mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}(t)|\right]}{\mathbb{E}\left[|\mathcal{R}(t)|\right]}. \tag{2}$$

In the sequel, we provide guarantees for both mFDR and FDR. Our guarantees on mFDR hold under the following weakening of (1). Define the filtration formed by the sequence of sigma-fields $\mathcal{F}^t := \sigma(R_1, \ldots, R_t)$, and let $\alpha_t := f_t(R_1, \ldots, R_{t-1})$, where f_t is an arbitrary function of the first $t-1$ indicators for rejection. Then, we say that the null p-values are *conditionally super-uniformly distributed* if the following holds:

$$\text{If the null hypothesis } H_t \text{ is true, then} \quad \Pr\{P_t \leq \alpha_t \mid \mathcal{F}^{t-1}\} \leq \alpha_t. \tag{3}$$

2.1. An Oracle FDP Estimate and a Naive Overestimate

To understand the motivation behind the new procedure, it is necessary to expand on an perspective on existing online FDR procedures, recently suggested by Ramdas et al. (2017). We begin by defining an *oracle estimate* of the FDP as:

$$\text{FDP}^*(t) := \frac{\sum\limits_{j \leq t, j \in \mathcal{H}^0} \alpha_j}{|\mathcal{R}(t)|}.$$

The word oracle indicates that FDP^* cannot be calculated by the scientist, since $\mathcal{H}^0$ is unknown. Intuitively, the numerator $\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j$ overestimates the number of false discoveries, and $\text{FDP}^*(t)$ overestimates the FDP, as formalized in the claim below:

Proposition 1. *If the null p-values are conditionally super-uniformly distributed* (3)*, then we have:*

(a) $\mathbb{E}\left[\sum\limits_{j \leq t, j \in \mathcal{H}^0} \alpha_j \right] \geq \mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}(t)| \right];$

(b) If $\text{FDP}^*(t) \leq \alpha$ *for all* $t \in \mathbb{N}$*, then* $\text{mFDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$*.*

Further, if the null p-values are independent of each other and of the non-nulls, and $\{\alpha_t\}$ *is a monotone function of past rejections, then:*

(c) $\mathbb{E}\left[\text{FDP}^*(t) \right] \geq \mathbb{E}\left[\text{FDP}(t) \right] \equiv \text{FDR}(t)$ *for all* $t \in \mathbb{N};$

(d) The condition $\text{FDP}^*(t) \leq \alpha$ *for all* $t \in \mathbb{N}$ *implies that* $\text{FDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$*.*

To clarify, the word *monotone* means that α_t is a coordinate-wise non-decreasing function of the vector $R_1, \ldots, R_{t-1}$. Proposition 1 follows from the results of Ramdas et al. (2017), and we prove it in Subsection 3.1 for completeness. Even though $\text{FDP}^*(t)$ cannot be directly calculated and used, Proposition 1 is a useful way to identify what would be ideally possible.

One natural way to convert $\text{FDP}^*(t)$ to a truly empirical overestimate of $\text{FDP}(t)$ is to define:

$$\widehat{\text{FDP}}_{\text{LORD}}(t) := \frac{\sum_{j \leq t} \alpha_j}{|\mathcal{R}(t)|}.$$

Since it is trivially true that $\widehat{\text{FDP}}_{\text{LORD}}(t) \geq \text{FDP}^*(t)$, we immediately obtain that Proposition 1 also holds with $\text{FDP}^*(t)$ replaced by $\widehat{\text{FDP}}_{\text{LORD}}(t)$. The subscript LORD is used because Ramdas et al. (2017) point out that their variant of the LORD algorithm of Javanmard and Montanari (2017) can be derived by simply assigning α_j in an online fashion to ensure that the condition $\widehat{\text{FDP}}_{\text{LORD}}(t) \leq \alpha$ is met for all times t.

2.2. A Better Estimate of the Alpha-Wealth Spent on Testing Nulls

The main drawback of $\widehat{\text{FDP}}_{\text{LORD}}$ is that if the underlying (unknown) truth is such that the proportion of non-nulls (true signals) is non-negligible, then $\widehat{\text{FDP}}_{\text{LORD}}(t)$ is a very crude and overly conservative overestimate of $\text{FDP}^*(t)$, and hence also of the true unknown FDP. With this drawback in mind, and knowing that we would expect non-nulls to typically have smaller p-values, we propose the following novel estimator:

$$\widehat{\text{FDP}}_{\text{SAFFRON}(\lambda)}(t) \equiv \widehat{\text{FDP}}_\lambda(t) := \frac{\sum_{j \leq t} \alpha_j \frac{\mathbf{1}\{P_j > \lambda_j\}}{1 - \lambda_j}}{|\mathcal{R}(t)|},$$

where $\{\lambda_j\}_{j=1}^\infty$ is a predictable sequence of user-chosen parameters in the interval $(0, 1)$. Here the term *predictable* means that λ_j is a deterministic function of the information available from time 1 to $j - 1$, which will be formalized later. For simplicity, when λ_j is chosen to be a constant for all j, we will drop the subscript and just write λ, and we will consider $\lambda = 1/2$ as our default choice. SAFFRON is based on the idea that the numerator of $\widehat{\text{FDP}}_\lambda$ is a much better estimator of the quantity $\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j$ than LORD's naive estimate $\sum_{j \leq t} \alpha_j$.

So as to provide some intuition for why we expect $\widehat{\text{FDP}}_\lambda$ to be a fairly tight estimate of FDP^*, note that $\frac{\mathbf{1}\{P_j > \lambda_j\}}{1 - \lambda_j}$ has a unit expectation whenever P_j is uniformly distributed (null), but would typically have a much smaller expectation whenever P_j is stochastically much smaller than uniform (non-null). The following theorem shows that, even though $\widehat{\text{FDP}}_\lambda(t)$ is not necessarily always larger than $\text{FDP}^*(t)$, a direct analog of Proposition 1 is nonetheless valid. In order to state this claim formally, we need to slightly modify the assumption (3). As before, denote by $R_j := \mathbf{1}\{P_j \leq \alpha_j\}$ the indicator for rejection, and let $C_j := \mathbf{1}\{P_j \leq \lambda_j\}$ be the indicator for candidacy. Accordingly, we refer to the p-values for which $C_j = 1$ as candidates. Moreover, we let $\alpha_t := f_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$, where f_t denotes an arbitrary function of the first $t - 1$ indicators for rejection and candidacy, and define the filtration generated from sigma-fields $\mathcal{F}^t := \sigma(R_1, \ldots, R_t, C_1, \ldots, C_t)$. With respect to this filtration, we introduce a conditional super-uniformity condition on the null p-values similar to (3):

$$\text{If the null hypothesis } H_t \text{ is true, then} \qquad (4)$$
$$\Pr\{P_t \leq \alpha_t \mid \mathcal{F}^{t-1}\} \leq \alpha_t,$$

which can be rephrased as:

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_t > \alpha_t\}}{1 - \alpha_t} \,\Big|\, \mathcal{F}^{t-1} \right] \geq 1 \geq \mathbb{E}\left[\frac{\mathbf{1}\{P_t \leq \alpha_t\}}{\alpha_t} \,\Big|\, \mathcal{F}^{t-1} \right].$$

Note that again marginal super-uniformity (1) implies this condition, provided that the p-values are independent.

Theorem 1. *If the null p-values are conditionally super-uniformly distributed* (4)*, then we have:*

(a) $\mathbb{E}\left[\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j \frac{\mathbf{1}\{P_j > \lambda_j\}}{1-\lambda_j}\right] \geq \mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}(t)|\right];$

(b) The condition $\widehat{\mathrm{FDP}}_\lambda(t) \leq \alpha$ *for all* $t \in \mathbb{N}$ *implies that* $\mathrm{mFDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$.

Further, if the null p-values are independent of each other and of the non-nulls, and $\{\alpha_t\}$ *is a monotone function of the vector* $R_1, ..., R_{t-1}, C_1, ..., C_{t-1}$*, then we additionally have:*

(c) $\mathbb{E}\left[\widehat{\mathrm{FDP}}_\lambda(t)\right] \geq \mathbb{E}\left[\mathrm{FDP}(t)\right] \equiv \mathrm{FDR}(t)$ *for all* $t \in \mathbb{N}$;

(d) The condition $\widehat{\mathrm{FDP}}_\lambda(t) \leq \alpha$ *for all* $t \in \mathbb{N}$ *implies that* $\mathrm{FDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$.

The proof of this theorem is given in Section 3.2, and is based on a "reverse super-uniformity lemma" that is discussed in the next section. This lemma, though of a technical nature, may be of independent interest in deriving new algorithms. The statements on mFDR control allow SAFFRON to be used in place of LORD in applications in which p-values are not independent, but are conditionally super-uniformly distributed, such as the MAB-FDR framework (based on multi-armed bandits) proposed by Yang et al. (2017).

2.3. The SAFFRON Algorithm for Constant λ

We now present the SAFFRON algorithm at a high level. For simplicity, we consider the constant λ setting, which performs well in experiments, though it may be a useful direction for future work to construct good heuristics for time-varying sequences $\{\lambda_j\}_{j=1}^\infty$.

1. Given a target FDR level α, the user first picks a constant $\lambda \in (0,1)$, an initial wealth $W_0 < (1-\lambda)\alpha$, and a positive non-increasing sequence $\{\gamma_j\}_{j=1}^\infty$ of summing to one. For example, given a parameter $s > 1$, we might pick $\gamma_j \propto j^{-s}$ for some $s > 1$.

2. We use the term "candidates" to refer to p-values smaller than λ, since SAFFRON will never reject a p-value larger than λ. Recalling the indicator for candidacy $C_t := \mathbf{1}\{P_t \leq \lambda\}$, and denoting by τ_j be the time of the j-th rejection (and setting $\tau_0 = 0$), define the candidates after the j-th rejection as $C_{j+} = C_{j+}(t) = \sum_{i=\tau_j+1}^{t-1} C_i$.

3. SAFFRON begins by allocation $\alpha_1 = \min\{\gamma_1 W_0, \lambda\}$,

and then at time $t = 2, 3, \ldots$, it allocates:

$$\alpha_t := \min\{\lambda, \tilde{\alpha}_t\}, \quad \text{where} \quad \tilde{\alpha}_t := W_0 \gamma_{t-C_{0+}} +$$

$$((1-\lambda)\alpha - W_0)\gamma_{t-\tau_1-C_{1+}} + \sum_{j \geq 2}(1-\lambda)\alpha\gamma_{t-\tau_j-C_{j+}}.$$

In words, SAFFRON starts off with an alpha-wealth $W_0 < (1-\lambda)\alpha$, never loses wealth when testing candidate p-values, gains wealth of $(1-\lambda)\alpha$ on every rejection except the first. If there is a significant fraction of non-nulls, and the signals are fairly strong, then SAFFRON may make more rejections than LORD.

To clarify, SAFFRON guarantees FDR control for any $\lambda \in (0,1)$ and any chosen sequence $\{\gamma_j\}_{j=1}^\infty$, but the algorithm's power, or ability to detect signals, varies as a function of these parameters. Given the minimal nature of our assumptions, there is no universally optimal constant or sequence: specifically, we do not make assumptions on the frequency of true signals, or on how strong they are, or on their order, all of which are factors that affect the power. We discuss reasonable default choices in the experimental section.

2.4. Relationship to Other Procedures

Here, we compare SAFFRON to existing procedures in the literature, emphasizing commonalities that allow us to give a unified view of seemingly disparate algorithms.

Alpha-investing. Even though the motivation that we have presented for SAFFRON relates it to the LORD algorithm, we find it interesting that the original alpha-investing algorithm of Foster and Stine (2008) is recovered by choosing $\lambda_j = \alpha_j$ in $\widehat{\mathrm{FDP}}_\lambda$, and attempting to ensure that $\widehat{\mathrm{FDP}}_\lambda(t) \leq \alpha$ for all times $t \in \mathbb{N}$. In order to see this fact, first note that with this choice of λ_j, the indicator $\mathbf{1}\{P_j > \lambda_j\}$ simply indicates when the j-th hypothesis is not rejected. Consequently, the numerator of $\widehat{\mathrm{FDP}}_\lambda$ reads as $\sum_{j \leq t} \frac{\alpha_j}{1-\alpha_j}\mathbf{1}\{j \notin \mathcal{R}(t)\}$. Hence, ensuring that $\widehat{\mathrm{FDP}}_\lambda(t) \leq \alpha$ at all times $t \in \mathbb{N}$, is equivalent to ensuring that $\sum_{j \leq t} \frac{\alpha_j}{1-\alpha_j}\mathbf{1}\{j \notin \mathcal{R}(t)\}$ never exceeds $\alpha(|\mathcal{R}(t)| \vee 1)$, which, in the language of alpha-investing, is equivalent to ensuring that the algorithm's wealth never becomes negative.[1] Just as Ramdas et al. (2017) were able to reinterpret and rederive LORD in terms of a particular estimate of the FDP, the current work allows us to reinterpret and rederive alpha-investing in terms of SAFFRON's FDP. However, our simulations demonstrate that despite this similarity, SAFFRON with $\lambda_j = 1/2$ is typically a more powerful algorithm than both LORD and alpha-investing.

---

[1] Recall that the alpha-investing algorithm starts off with an alpha-wealth of α, reduces its alpha-wealth by $\frac{\alpha_j}{1-\alpha_j}$ after tests that fail to reject, and increase the wealth by α on rejections.

Storey-BH. In offline multiple testing, where all n p-values are immediately available, the Benjamini-Hochberg (BH) procedure (1995) is a classical method for guaranteeing FDR control. BH estimates the FDP of the rejection set $\mathcal{R}(s) := \{i : P_i \leq s\}$ by $\widehat{\text{FDP}}_{\text{BH}}(s) := \frac{n \cdot s}{|\mathcal{R}(s)|}$, which is a conservative estimate of the oracle $\text{FDP}^*_{\text{BH}}(s) := \frac{|\mathcal{H}^0| \cdot s}{|\mathcal{R}(s)|}$ (details in Supplementary Material). For independent p-values, Storey et al. (2002; 2004) improved the BH method, by picking a constant $\lambda \in (0, 1)$, and calculating:

$$\widehat{\text{FDP}}_{\text{StBH}}(s) := \frac{n \cdot s \cdot \widehat{\pi_0}}{|\mathcal{R}(s)|},$$

where $\widehat{\pi_0}$ is an estimate of the unknown proportion of nulls $\pi_0 = |\mathcal{H}^0|/n$ computed as:

$$\widehat{\pi_0} := \frac{1 + \sum_{i=1}^n \mathbf{1}\{P_i > \lambda\}}{n(1 - \lambda)}.$$

Then, this procedure, which we refer to as "Storey-BH," calculates $\widehat{s}_{\text{StBH}} := \max\{s : \widehat{\text{FDP}}_{\text{StBH}}(s) \leq \alpha\}$ and rejects the set $\mathcal{R}(\widehat{s}_{\text{StBH}})$ which satisfies the bound FDR $\leq \alpha$. Procedures such as Storey-BH are known in the multiple testing literature as *adaptive procedures*, since they automatically adapt to the unknown proportion of nulls.

Returning to the setting of online FDR, what matters is not the proportion of nulls π_0, but instead a running estimate of the amount of alpha-wealth that was spent testing nulls thus far; this difference arises because, unlike the offline setting where all p-values are compared to the same level $\widehat{s}$, different p-values have to pass different thresholds α_i. In light of the above discussion, it should be apparent that Storey-BH is to BH as SAFFRON is to LORD. In the offline context, Storey-BH is called an "adaptive method" (it is adaptive to the unknown null proportion) and in this sense, SAFFRON can be seen as an adaptive online FDR method.

Accumulation tests. Note that $\mathbb{E}[2I(P > 1/2)] \geq 1$ for null p-values (with equality when they are exactly uniformly distributed, simply because $\int_0^1 2I(p > 1/2)dp = 1$). One may actually use any non-decreasing function h such that $\int_0^1 h(p)dp$ in the formula for $\widehat{\text{FDP}}_\lambda$. Such *accumulation functions* were studied (Li & Barber, 2017) in the (offline) context of ordered testing, and may seamlessly be transferred to the online setting considered here, yielding mFDR control using the same proof. In initial experiments, the use of other functions does not seem to yield any advantage, and under some additional assumptions in the offline ordered testing setting, the aforementioned authors argued that the step function $(1 - \lambda)^{-1}I(I > \lambda)$ is asymptotically optimal for power. In this light, SAFFRON can also be seen as an online analog of adaptive SeqStep (Lei & Fithian, 2016), which is a variant of selective SeqStep (Barber & Candès, 2015) and SeqStep (Li & Barber, 2017).

3. Proof of Theorem 1 Using a Reverse Super-Uniformity Lemma

In this section, we present a lemma that is central to the proof of FDR control for SAFFRON. We then use this lemma to prove Proposition 1 and Theorem 1. Let us first recall and set up some preliminary notation. In what follows, α_t, λ_t are random variables in $(0, 1)$ that always satisfy $\alpha_t \leq \lambda_t$. We denote the indicator for rejection at the t-th step by $R_t := \mathbf{1}\{P_t \leq \alpha_t\}$, and recall that since only p-values smaller than λ_t are candidates for rejection, we had earlier defined the indicator for candidacy as $C_t := \mathbf{1}\{P_t \leq \lambda_t\}$. If we denote $\bar{C}_t = 1 - C_t$, then it is clear that $R_t \bar{C}_t = 0$, since $R_t = 1$ implies $\bar{C}_t = 0$ and $\bar{C}_t = 1$ implies $R_t = 0$, and it is also possible for R_t and $\bar{C}_t$ to both equal 0. Also let $R_{1:t} := \{R_1, \ldots, R_t\}$ and $C_{1:t} := \{C_1, \ldots, C_t\}$. As before, we consider the filtration $\mathcal{F}^t := \sigma(R_{1:t}, C_{1:t})$. In what follows, we insist that the sequences $\{\alpha_t\}_{t=1}^\infty$ and $\{\lambda_t\}_{t=1}^\infty$ are *predictable*, meaning that they are functions of the information available from time 1 to $t - 1$ only; specifically, we insist that α_t, λ_t are measurable with respect to the sigma-field $\mathcal{F}^{t-1}$. We will also require that the $\{\alpha_t\}$ sequence is *monotone*, meaning that $\alpha_t = f_t(R_{1:t-1}, C_{1:t-1})$ for some coordinatewise non-decreasing function $f_t : \{0, 1\}^{2(t-1)} \to [0, \lambda_t]$. The proof that SAFFRON as described in Subsection 2.3 satisfies this requirement is given in the Supplementary Material.

Recall the definition (4) of conditional super-uniformity, as well as its equivalent rephrased form in the line after definition (4). Lemma 1 guarantees that for independent p-values, this statement holds true more generally.

Lemma 1. *Assume that the p-values $P_1, P_2, \ldots$ are independent and let $g : \{0, 1\}^T \to \mathbb{R}$ be any coordinatewise non-decreasing function. Then, for any index $t \leq T$ such that $H_t \in \mathcal{H}^0$, we have:*

$$\mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})\mathbf{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)g(R_{1:T})} \,\middle|\, \mathcal{F}^{t-1}\right]$$

$$\geq \mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})}{g(R_{1:T})} \,\middle|\, \mathcal{F}^{t-1}\right]$$

$$\geq \mathbb{E}\left[\frac{\mathbf{1}\{P_t \leq f_t(R_{1:t-1}, C_{1:t-1})\}}{g(R_{1:T})} \,\middle|\, \mathcal{F}^{t-1}\right].$$

Proof. The second inequality is a consequence of super-uniformity lemmas from past work (Ramdas et al., 2017; Javanmard & Montanari, 2017), so we only prove the first inequality. At a high level, the proof strategy is inverted, and we will hallucinate a vector with one element being set to 1, instead of being set to 0 in the aforementioned works.

Letting $P_{1:T} = (P_1, \ldots, P_T)$ be the original vector of p-values, we define a "hallucinated" vector of p-values $\widetilde{P}_{1:T}^{t \to 1} := (\widetilde{P}_1, \ldots, \widetilde{P}_T)$ that equals $P_{1:T}$, except that the

t-th component is set to one:

$$\widetilde{P}_i = \begin{cases} 1 & \text{if } i = t \\ P_i & \text{if } i \neq t. \end{cases}$$

Define hallucinated candidate and rejection indicators as $\widetilde{C}_i = \mathbf{1}\left\{\widetilde{P}_i \leq \lambda_i\right\}$ and $\widetilde{R}_i = \mathbf{1}\left\{\widetilde{P}_i \leq f_i(\widetilde{R}_{1:i-1}, \widetilde{C}_{1:i-1})\right\}$ respectively. Let $R_{1:T} = (R_1, \ldots, R_T)$ and $\widetilde{R}_{1:T}^{t \to 1} = (\widetilde{R}_1, \ldots, \widetilde{R}_T)$ denote the vector of rejections using $P_{1:T}$ and $\widetilde{P}_{1:T}^{t \to 1}$, respectively. Similarly, let $C_{1:T} = (C_1, \ldots, C_T)$ and $\widetilde{C}_{1:T}^{t \to 1} = (\widetilde{C}_1, \ldots, \widetilde{C}_T)$ denote the vector of candidates using $P_{1:T}$ and $\widetilde{P}_{1:T}^{t \to 1}$, respectively.

By construction, we have the following properties:

1. $\widetilde{R}_i = R_i$ and $\widetilde{C}_i = C_i$ for all $i < t$, hence $f_i(R_{1:i-1}, C_{1:i-1}) = f_i(\widetilde{R}_{1:i-1}, \widetilde{C}_{1:i-1})$ for all $i \leq t$.

2. $\widetilde{R}_t = \widetilde{C}_t = 0$, and hence $\widetilde{R}_i \leq R_i$ for all $i \geq t$, due to monotonicity of the functions f_i.

Hence, on the event $\{P_t > \lambda_t\}$, we have $R_t = \widetilde{R}_t = 0$ and $C_t = \widetilde{C}_t = 0$, and hence also $R_{1:T} = \widetilde{R}_{1:T}^{t \to 1}$. This allows us to conclude that:

$$\frac{f_t(R_{1:t-1}, C_{1:t-1})\mathbf{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)g(R_{1:T})} = \frac{f_t(R_{1:t-1}, C_{1:t-1})\mathbf{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)g(\widetilde{R}_{1:T}^{t \to 1})}.$$

Since $\widetilde{R}_{1:T}^{t \to 1}$ is independent of P_t, we may take conditional expectations to obtain:

$$\mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})\mathbf{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)g(R_{1:T})} \,\middle|\, \mathcal{F}^{t-1}\right]$$

$$= \mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})\mathbf{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)g(\widetilde{R}_{1:T}^{t \to 1})} \,\middle|\, \mathcal{F}^{t-1}\right]$$

$$\overset{(i)}{\geq} \mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})}{g(\widetilde{R}_{1:T}^{t \to 1})} \,\middle|\, \mathcal{F}^{t-1}\right]$$

$$\overset{(ii)}{\geq} \mathbb{E}\left[\frac{f_t(R_{1:t-1}, C_{1:t-1})}{g(R_{1:T})} \,\middle|\, \mathcal{F}^{t-1}\right],$$

where inequality (i) follows by taking an expectation only with respect to P_t by invoking the conditional super-uniformity property (4); and inequality (ii) follows because $g(R_{1:T}) \geq g(\widetilde{R}_{1:T}^{t \to 1})$ since $R_i \geq \widetilde{R}_i$ for all i by monotonicity of the online FDR rule. This concludes the proof of the lemma. $\square$

We now proceed to using the above lemma to prove Proposition 1 and Theorem 1.

3.1. Proof of Proposition 1

Statement (a) is proved by noting that for any time $t \in \mathbb{N}$, we have:

$$\mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}(t)|\right] = \sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{E}\left[\mathbf{1}\{P_j \leq \alpha_j\}\right]$$

$$\leq \sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{E}\left[\alpha_j\right],$$

where the inequality follows after taking iterated expectations by conditioning on $\mathcal{F}^{j-1}$, and then applying the conditional super-uniformity property (3).

If we have $\text{FDP}^*(t) := \frac{1}{|\mathcal{R}(t)|} \sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j \leq \alpha$, as assumed in statement (b), then it follows that:

$$\sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{E}\left[\alpha_j\right] = \mathbb{E}\left[\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j\right]$$

$$\leq \alpha\mathbb{E}\left[|\mathcal{R}(t)|\right],$$

using linearity of expectation and the assumption on $\text{FDP}^*(t)$. Using part (a) and rearranging yields the inequality $\text{mFDR}(t) := \frac{\mathbb{E}[|\mathcal{H}^0 \cap \mathcal{R}(t)|]}{\mathbb{E}[|\mathcal{R}(t)|]} \leq \alpha$, which concludes the proof of part (b).

If, in addition, the null p-values are independent of each other and of the non-nulls and the sequence $\{\alpha_t\}$ is monotone, we can use the following argument to prove claims (c) and (d). These claims establish that the procedure controls the FDR at any time $t \in \mathbb{N}$. Still assuming the inequality $\text{FDP}^*(t) \leq \alpha$, we have:

$$\text{FDR}(t) = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}(t)|}{|\mathcal{R}(t)|}\right]$$

$$= \sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\mathbf{1}\{P_j \leq \alpha_j\}}{|\mathcal{R}(t)|}\right]$$

$$\leq \sum_{j \leq t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_j}{|\mathcal{R}(t)|}\right]$$

$$= \mathbb{E}\left[\text{FDP}^*(t)\right]$$

$$\leq \alpha,$$

where the first inequality follows after taking iterated expectations by conditioning on $\mathcal{F}^{j-1}$, and then applying the super-uniformity lemma (Ramdas et al., 2017), the following equality uses linearity of expectation, and the final inequality follows by the assumption on $\text{FDP}^*(t)$. This concludes the proof of both statements (c) and (d).

3.2. Proof of Theorem 1

First note that, for any time $t \in \mathbb{N}$, we have:

$$\mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}(t)|\right] = \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\mathbf{1}\left\{P_j \le \alpha_j\right\}\right]$$

$$\overset{(i)}{\le} \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\alpha_j\right]$$

$$\overset{(ii)}{\le} \mathbb{E}\left[\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j \frac{\mathbf{1}\left\{P_j > \lambda_j\right\}}{1 - \lambda_j}\right],$$

where inequality (i) first uses the law of iterated expectations by conditioning on $\mathcal{F}^{j-1}$, and then both (i) and (ii) apply the conditional super-uniformity property (4), which concludes the proof of part (a). To prove part (b), we drop the condition $j \in \mathcal{H}^0$ from the last expectation, and use the assumption that $\widehat{\mathrm{FDP}}_\lambda(t) := \frac{\sum_{j \le t} \alpha_j \frac{\mathbf{1}\{P_j > \lambda_j\}}{1 - \lambda_j}}{|\mathcal{R}(t)|} \le \alpha$ to obtain:

$$\mathbb{E}\left[\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j \frac{\mathbf{1}\left\{P_j > \lambda_j\right\}}{1 - \lambda_j}\right] \le \alpha \mathbb{E}\left[|\mathcal{R}(t)|\right].$$

Combining this inequality with the result of part (a), and rearranging the terms, we reach the conclusion that $\mathrm{mFDR}(t) \le \alpha$, as desired. Under the independence and monotonicity assumptions of parts (c, d), we have

$$\mathrm{FDR}(t) = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}(t)|}{|\mathcal{R}(t)|}\right]$$

$$= \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_j \le \alpha_j\right\}}{|\mathcal{R}(t)|}\right]$$

$$\overset{(iii)}{\le} \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_j}{|\mathcal{R}(t)|}\right]$$

$$\overset{(iv)}{\le} \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_j \mathbf{1}\left\{P_j > \lambda_j\right\}}{(1 - \lambda_j)|\mathcal{R}(t)|}\right],$$

where inequality (iii) first uses iterated expectations by conditioning on $\mathcal{F}^{j-1}$, and then both (iii) and (iv) apply Lemma 1. Assuming that the inequality $\widehat{\mathrm{FDP}}_\lambda(t) \le \alpha$ holds, it follows that:

$$\sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_j \mathbf{1}\left\{P_j > \lambda_j\right\}}{(1 - \lambda_j)|\mathcal{R}(t)|}\right] \overset{(v)}{\le} \mathbb{E}\left[\frac{\sum_{j \le t} \alpha_j \mathbf{1}\left\{P_j > \lambda_j\right\}}{(1 - \lambda_j)|\mathcal{R}(t)|}\right]$$

$$\overset{(vi)}{=} \mathbb{E}\left[\widehat{\mathrm{FDP}}_\lambda(t)\right]$$

$$\overset{(vii)}{\le} \alpha,$$

where inequality (v) follows by linearity of expectation and summing over a larger set of indices; equality (vi) simply uses the definition of $\widehat{\mathrm{FDP}}_\lambda(t)$, and inequality (vii) follows by the assumption, hence proving parts (c,d).

4. Numerical Simulations

In this section, we provide the results of some numerical experiments that compare the performance of SAFFRON with current state-of-the-art algorithms for online FDR control, namely the aforementioned LORD and alpha-investing procedures.[2] For each method, we provide empirical evaluations of its power while ensuring that the FDR remains below a chosen value. We only run simulations since for real data, we would not know the ground truth and hence which discoveries are true or false.

The following two subsections separately analyze two experimental settings - one in which the p-values are computed from Gaussian observations, and another in which the p-values under the alternative are drawn from a beta distribution. In both cases, SAFFRON outperforms the competing algorithms, with the exact level of performance depending on the choice of sequence $\{\gamma_j\}$. All experiments use a target FDR of $\alpha = 0.05$ and estimate the FDR and power by averaging over 200 independent trials. As previously mentioned, the constant sequence $\lambda_j = 1/2$ for all j was found to be particularly successful, so this is our default choice, and we drop the index for simplicity.

4.1. Testing with Gaussian Observations

We use the simple experimental setup of testing the mean of a Gaussian distribution with $T = 1000$ components. More precisely, for each index $i \in \{1, \dots, T\}$, the null hypothesis takes the form $H_i : \mu_i = 0$. The observations consist of independent Gaussian variates $Z_i \sim N(\mu_i, 1)$, which are converted into one-sided p-values using the transform $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian CDF. The motivation for one-sided conversion lies in A/B testing, where one wishes to detect *larger* effects, not smaller. The parameter μ_i is chosen according to a mixture model:

$$\mu_i = \begin{cases} 0 & \text{with probability } 1 - \pi_1 \\ F_1 & \text{with probability } \pi_1, \end{cases}$$

where the random variable F_1 is of the form $N(\mu_c, 1)$ for some constant μ_c. We ran simulations for $\mu_c \in \{2, 3\}$, thus seeing how changing the distance of the alternative mean to the null mean affects the performance of SAFFRON.

In what follows, we compare SAFFRON's achieved power and FDR to those of LORD and alpha-investing. The constant infinite sequence $\gamma_j \propto \frac{\log(j \vee 2)}{j e^{\sqrt{\log j}}}$, where the proportionality constant is determined so that the sequence sums to one, was shown to be asymptotically optimal for testing Gaussian means via the LORD method in the paper (Javanmard & Montanari, 2017). Since SAFFRON loses wealth only when

---

[2]The code for all simulations described in this section is available at: https://github.com/tijana-zrnic/SAFFRONcode

testing non-candidates whereas LORD loses wealth at every step, it is expected to behave more conservatively and not use up its wealth at the same rate, conditioned on both using the same sequence $\{\gamma_j\}$. For this reason, informally speaking, it can reuse this leftover wealth, hence the sequence $\{\gamma_j\}$ chosen for SAFFRON is more aggressive, in the sense that more wealth is concentrated around the beginning of the sequence. In particular, we choose sequences of the form $\gamma_j \propto j^{-s}$, where the parameter $s > 1$ controls the aggressiveness of the procedure; the greater the constant s, the more wealth is concentrated around small values of j. We also consider these sequences for LORD, thus observing the difference in performance resulting from using a more aggressive sequence in the regime of a finite sequence of p-values. Figures showing the power and FDR of SAFFRON and LORD by varying the aggressiveness of sequence $\{\gamma_j\}$ are in the Supplementary Material.

In Figure 2 we consider $F_1 = N(2, 1)$, and compare the level of performance of alpha-investing, SAFFRON and LORD, the latter two using the highest performing sequence chosen among six possible sequences. Figure 1 demonstrates the same comparison for a similar but somewhat easier testing problem, with $F_1 = N(3, 1)$. Experiments indicate that increasing the fraction of non-null hypotheses allows SAFFRON to achieve a faster increase of power than LORD, thus performing considerably better than both LORD and the alpha-investing procedure in settings with a great number of non-null observations.
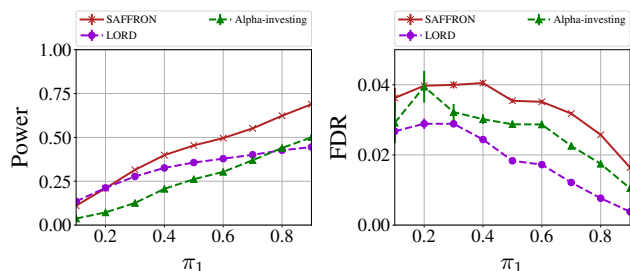

Figure 2. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON, LORD and alpha-investing, the first two using the sequence $\{\gamma_j\}$ which achieves the highest power for each of them (chosen over six sequences of varying aggressiveness). The observations under the alternative are Gaussian with $\mu_i \sim N(2, 1)$ and standard deviation 1, and are converted into one-sided p-values as $P_i = \Phi(-Z_i)$.

4.2. Testing with Beta Alternatives

In this setting we generate the p-value sequence according to the following model:

$$P_i \sim \begin{cases} \text{Unif}[0, 1], & \text{with probability } 1 - \pi_1 \\ \text{Beta}(m, n), & \text{with probability } \pi_1, \end{cases}$$

where $i \in [T]$ and $T = 1000$, as before. Again we compare the performance of SAFFRON, alpha-investing and LORD in terms of the achieved power with the FDR controlled under a chosen level. For LORD, the asymptotically optimal sequence $\{\gamma_j\}$ was derived in the paper (Javanmard & Montanari, 2017) and is of the form $\gamma_j \propto (\frac{1}{j} \log j)^{1/m}$ for $m < 1$ and $n \geq 1$. As in the Gaussian case, for SAFFRON and additionally for LORD we consider the sequence $\gamma_j \propto j^{-s}$ with varying s, which, unlike the aforementioned sequence, does not depend on the parameters of the distribution. Please refer to the Supplementary Material for plots of achieved power and FDR of SAFFRON and LORD obtained by varying the sequence. For the particular distribution of the observed p-values we choose $m = 0.5$ and $n = 5$.

Figure 3 compares the performance of SAFFRON, LORD and alpha-investing, the first two using the highest performing sequence $\{\gamma_j\}$ chosen among six considered sequences, as in the setting with Gaussian tests. Although simulations show SAFFRON performing similarly to LORD and alpha-investing for small fractions of non-null hypotheses, it significantly outperforms its competitors in terms of power and using up available wealth with a higher number of signals.
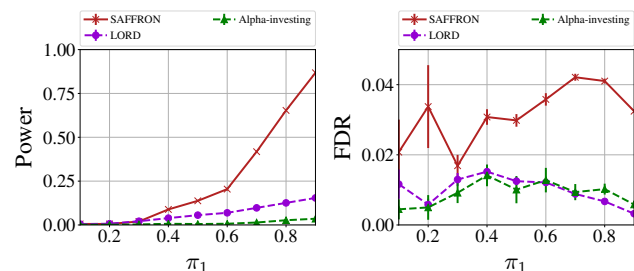

Figure 3. Statistical power and FDR versus fraction of non-null hypotheses π_1 for SAFFRON, LORD and alpha-investing, the first two using the sequence $\{\gamma_j\}$ which achieves the highest power for each of them (chosen over six sequences of varying aggressiveness). Alternative p-values are distributed as $\text{Beta}(0.5, 5)$.

5. Summary

This paper introduces SAFFRON, a new algorithmic framework for online mFDR and FDR control. We show empirically that SAFFRON is more powerful than existing algorithms. SAFFRON is based on a novel reverse superuniformity lemma that allows us to estimate the fraction of alpha-wealth that an algorithm spends on testing null hypotheses. One may interpret SAFFRON as an adaptive version of LORD, just as Storey-BH is an adaptive version of the Benjamini-Hochberg algorithm. Lastly, the derivation of SAFFRON is rather different from that of earlier generalized alpha-investing (GAI) algorithms, and as such provides a template for the derivation of new algorithms.

References

Aharoni, E. and Rosset, S. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 76(4):771–794, 2014.

Barber, R. F. and Candès, E. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

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

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

Javanmard, A. and Montanari, A. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, to appear, 2017.

Lei, L. and Fithian, W. Power of ordered hypothesis testing. In *International Conference on Machine Learning*, pp. 2924–2932, 2016.

Li, A. and Barber, R. F. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017.

Ramdas, A., Yang, F., Wainwright, M., and Jordan, M. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*, pp. 5655–5664, 2017.

Storey, J. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Storey, J., Taylor, J., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Tukey, J. *The Problem of Multiple Comparisons: Introduction and Parts A, B, and C*. Princeton University, 1953.

Yang, F., Ramdas, A., Jamieson, K., and Wainwright, M. J. Multi-A(rmed)/B(andit) testing with online FDR control. *Advances in Neural Information Processing Systems*, pp. 5959–5968, 2017.