Journal of Machine Learning Research 22 (2021) 1-39 Submitted 11/19; Revised 1/21; Published 1/21

Asynchronous Online Testing of Multiple Hypotheses

Tijana Zrnic TIJANA.ZRNIC@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720, USA

Aaditya Ramdas ARAMDAS@CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Editor: Gabor Lugosi

Abstract

We consider the problem of *asynchronous online testing*, aimed at providing control of the false discovery rate (FDR) during a continual stream of data collection and testing, where each test may be a sequential test that can start and stop at arbitrary times. This setting increasingly characterizes real-world applications in science and industry, where teams of researchers across large organizations may conduct tests of hypotheses in a decentralized manner. The overlap in time and space also tends to induce dependencies among test statistics, a challenge for classical methodology, which either assumes (overly optimistically) independence or (overly pessimistically) arbitrary dependence between test statistics. We present a general framework that addresses both of these issues via a unified computational abstraction that we refer to as "conflict sets." We show how this framework yields algorithms with formal FDR guarantees under a more intermediate, local notion of dependence. We illustrate our algorithms in simulations by comparing to existing algorithms for online FDR control.

Keywords: FDR control, false discovery rate, sequential hypothesis testing, sequential experimentation, p-values

1. Introduction

As applications of machine learning expand in scope beyond the classical setting of a single decision-maker and a single dataset, the decision-making side of the field has become increasingly important. Unfortunately, research on the decision-making side of the field has lagged relative to the pattern-recognition side, often focusing only on the validity of single decisions. Arguably, however, deployed machine learning models witness large collections of decisions, typically occurring in an extended asynchronous stream. In such settings, it is essential to consider error rates over sets of decisions, and not merely over single decisions.

Although it is not a focus of research in machine learning, multiple decision-making has been prominent during the past two decades in statistics, in the wake of seminal research by Benjamini and Hochberg (1995) on *false discovery rate* (FDR) control in multiple testing. That literature has, however, principally focused on batch data analysis and relatively small-scale problems. Modern applications in domains such as medicine, commerce, finance, and transportation are increasingly of planetary scale, with statistical analysis and decision-making tools being used to evaluate hundreds or thousands of related hypotheses in small windows of time (see, e.g., Tang et al., 2010; Xu et al., 2015). These testing processes are often sequential,

conducted in the context of a continuing stream of data analysis. The sequentiality is at two levels—each individual test is often a sequential procedure, terminating at a random time when a stopping criterion is satisfied, and also the overall set of tests is carried out sequentially, with possible overlap in time. In this setting—which we refer to as *asynchronous online testing*—the goal is to control a criterion such as the FDR not merely at the end of a batch of tests, but at any moment in time, and to do so while recognizing that the decision for a given test must generally be made while other tests are ongoing.

A recent literature on "online FDR control" has responded to one aspect of this problem, namely the problem of providing FDR control during a sequence of tests, and not merely at the end, by adaptively setting the test levels for the tests (Foster and Stine, 2008; Javanmard and Montanari, 2018; Ramdas et al., 2017, 2018). These methods are *synchronous*, meaning that each test can only start when the previous test has finished. Our goal is to consider the more realistic setting in which each test is itself a sequential process and where tests can overlap in time. This is done in real applications to gain time efficiency, and because of the difficulties of coordination in a large-scale, decentralized setting. To illustrate this point, Figure 1 compares the testing of five hypotheses within an asychronous setting and a synchronous setting. In the asynchronous setting, the test level α_t used to test hypothesis H_t is allowed to depend only on the outcomes of the previously *completed* tests—for example, α_3 can depend on the outcome of H_1, however not on the outcome of H_2. In the synchronous setting, on the other hand, the test level α_t can depend on all previously started (hence also completed) tests. To account for the uncertainty about the tests in progress, the test levels assigned by asynchronous online procedures must be more conservative. Thus, there is a tradeoff—although asynchronous procedures take less time to perform a given number of tests they are necessarily less powerful than their synchronous counterparts. The management of this tradeoff involves consideration of the overall power achieved per unit of real time, and consideration of the complexity of the coordination required in the synchronous setting.

Another limitation of existing work on online multiple testing is that the dependence assumptions on the tested sequence of test statistics, under which the formal false discovery rate guarantees hold, are usually at one of two extremes—they are either assumed to be independent, or arbitrarily dependent. From a practical perspective, independence seems overly optimistic as new tests may use previously collected data to formulate hypotheses, or to form a prior, or as evidence while testing. On the other hand, arbitrary dependence is likely too pessimistic, as older data and test outcomes with time become "stale," and no longer have direct influence on newly created tests. We see that a reconsideration of dependence is natural in the setting of online FDR control, and is particularly natural in the asynchronous setting, given that tests that are being conducted concurrently are often likely to be dependent, since they may use the same or highly correlated data during their overlap.

We therefore define and study a notion of *local dependence*, and place it within the context of asynchronous multiple testing. Working with p-values for simplicity, and letting P_t denote the t-th tested p-value, we say that a sequence of p-values $\{P_t\}$ satisfies local dependence if the following condition holds:

$$\text{for all } t > 0, \text{ there exists } L_t \in \mathbb{N} \text{ such that } P_t \perp P_{t-L_t-1}, P_{t-L_t-2}, \ldots, P_1, \tag{1}$$

where $\{L_t\}$ is a fixed sequence of parameters which we will refer to as "lags." Clearly, when $L_t = 0$ for all t, we obtain the independent setting, and when $L_t = t$, we recover the arbitrarily dependent setting. If $L_t \equiv L$ for all t, condition (1) captures a lagged dependence of order L.

To further emphasize the natural connection between asynchrony and local dependence, consider the simple setting in Figure 2. This diagram captures the setting in which a research team is collecting data over time, and decides to run multiple tests in a relatively short time interval. For example, such a situation might arise when testing multiple treatments against a common control (Robertson and Wason, 2018), or in large-scale A/B testing by internet companies (Xu et al., 2015). Since there is overlap in the data these tests use to compute their test statistics, the corresponding p-values could be arbitrarily dependent. In general several tests might share data with the first test. Thus the p-values are locally dependent, with the lag parameter being equal to the number of consecutive tests that share data.

In this work, we reinforce this connection between asynchronous online testing and dependence by developing a general abstract framework in which, from an algorithmic point of view, these two issues are treated

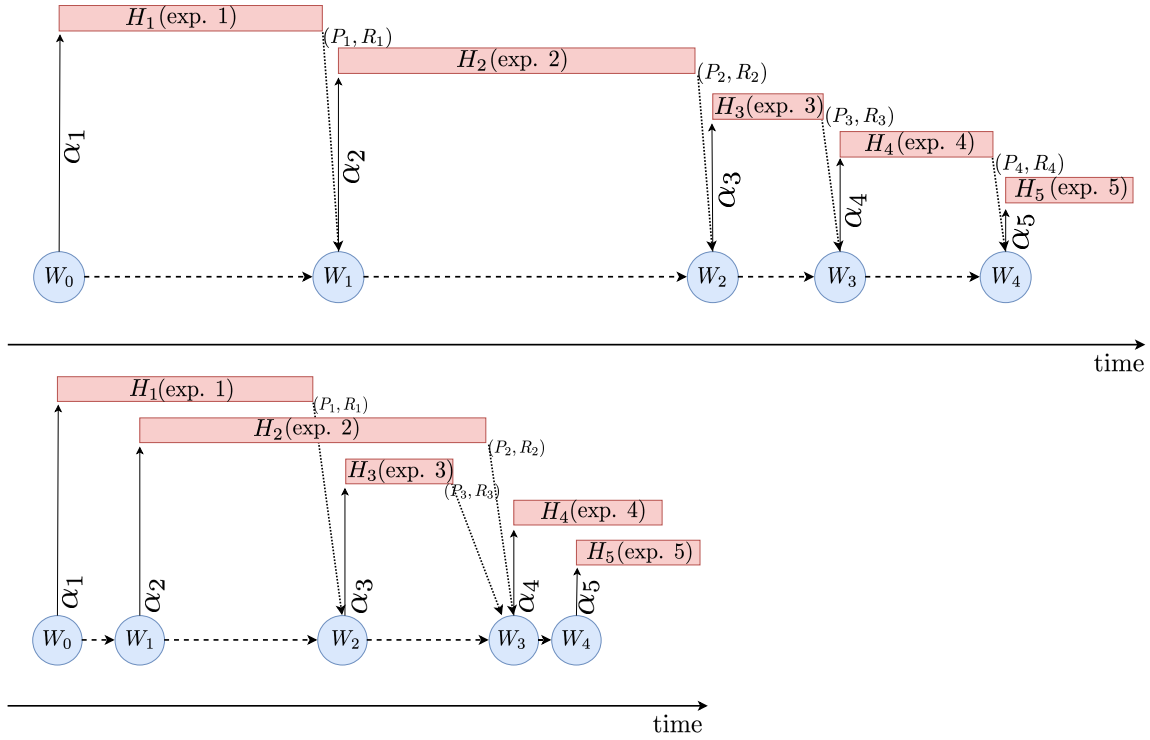

Figure 1: Testing five hypotheses synchronously (top) and asynchronously (bottom). In both cases, the test levels α_t depend on the outcomes of previously completed tests, which in the synchronous case includes all previously started tests. At the start time of experiment t, W_{t-1} is used to denote the remaining "wealth" for making false discoveries. At the end of experiment t, a p-value P_t and its corresponding decision $R_t := \mathbf{1}\{P_t \leq \alpha_t\}$ are known, which is used to adjust the available wealth at the start time of the next new test.

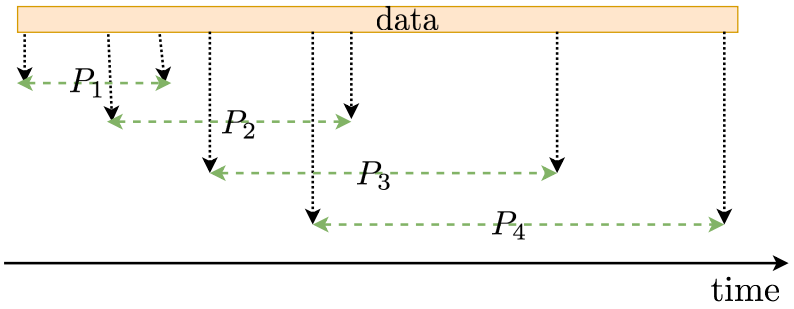

Figure 2: Example of p-values within a short interval computed on overlapping data. They exhibit local dependence; for example, P_3 and P_4 are independent of P_1.

with a single formal structure. We do so by associating with each test a *conflict set*, which consists of other tests that have a potentially adversarial relationship with the test in question. Within this framework, we develop algorithms with provable guarantees on the rate of false discoveries. The core idea is to enforce a notion of pessimism with regard to the conflict set—when computing a new test level, the algorithm "hallucinates" the worst-case outcomes of the conflicting tests.

We derive procedures that handle conflict sets as strict generalizations of current state-of-the-art online FDR procedures; indeed, when there are no conflicts, for example when there is no asynchrony and when the

p-values are independent, our solutions recover LORD (Javanmard and Montanari, 2018), LOND (Javanmard and Montanari, 2015), and SAFFRON (Ramdas et al., 2018), the latter of which recovers alpha-investing (Foster and Stine, 2008) as a special case for a particular choice of parameters. On the other hand, if the conflict sets are as large as possible—for example, if the parameter L_t or the number of tests run in parallel tend to infinity—our algorithms behave like alpha-spending,[1] which was designed to control a more stringent criterion called the family-wise error rate (FWER), under any dependence structure. Independently, we also prove that the original LOND procedure controls the FDR even under positive dependence (PRDS), the first online procedure to provably have this guarantee under the PRDS condition that is popular in the offline FDR literature (Benjamini and Yekutieli, 2001; Ramdas et al., 2019).

Organization. The rest of this paper is organized as follows. After a presentation of the general problem formulation and related work, Section 2 presents the key notion of conflict sets. We present two general procedures based on conflict sets, deferring their formal FDR guarantees to Section 6. In Section 3, we couch asynchronous testing in terms of conflict sets. In a similar fashion, in Section 4, we describe synchronous testing of locally dependent p-values using conflict sets, and present procedures having FDR guarantees within this environment. Section 5 then combines the ideas of local dependence and asynchronous testing into an overall framework designed for testing asynchronous batches of dependent p-values. Section 6 provides additional, stronger guarantees of the presented algorithms, which hold under more stringent assumptions on the p-value sequence. In Section 7 we present simulations designed to explore our methods, comparing them to existing procedures that handle dependent p-values. Finally, we conclude the paper with a short discussion in Section 8. All proofs are deferred to the Appendix.

1.1 Technical preliminaries

We briefly overview the technical background upon which our work builds. Recall that the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995) is defined as follows:

$$\text{FDR} \equiv \mathbb{E}\left[\text{FDP}\right] = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}|}{|\mathcal{R}| \vee 1}\right],$$

where $\mathcal{H}^0$ is the unknown set of true null hypotheses and $\mathcal{R}$ is the set of hypotheses rejected by some procedure. Formally we have $\mathcal{H}^0 = \{i : H_i \text{ is true}\}$, $\mathcal{R} = \{i : H_i \text{ is rejected}\}$. The random ratio appearing inside the expectation is called the *false discovery proportion* (FDP). It is also of theoretical and practical interest to study a related metric called the *modified false discovery rate* (mFDR):

$$\text{mFDR} \equiv \frac{\mathbb{E}\left[|\mathcal{H}^0 \cap \mathcal{R}|\right]}{\mathbb{E}\left[|\mathcal{R}| \vee 1\right]}.$$

Foster and Stine (2008) show that in the long run the mFDR behaves similarly to the FDR in an online environment. Similarly, Genovese and Wasserman (2002) prove that the mFDR and FDR achieved by the celebrated Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) become equivalent as the number of hypotheses tends to infinity. In this work, we mainly focus on the control of mFDR, as we can provide simple proofs under less restrictive assumptions. Importantly, in the Appendix we provide a side-by-side comparison of the mFDR and FDR for all of the experiments in this paper; as we show there, the plots for mFDR and FDR are visually indistinguishable when the number of non-nulls is non-negligible, and mFDR dominates the FDR when non-nulls are sparse. Thus, our experiments suggest that mFDR control suffices for FDR control as well.

In addition, we point out one advantage of mFDR over FDR which is especially relevant in the online context. Suppose that different sequences of hypotheses are tested with different algorithms controlling the mFDR. Then, one can retroactively group the set of discoveries resulting from these different algorithms,

1. Alpha-spending is a generalization of the Bonferroni correction in which the assigned test levels do not have to be equal. In other words, the Bonferroni correction suggests testing n hypotheses under level α/n, while alpha-spending merely requires $\sum_{i=1}^{n} \alpha_i \leq \alpha$, where α_i is the test level for the i-th hypothesis.

all the while knowing that the mFDR is still controlled. If the original sets of discoveries come with FDR guarantees only, one cannot argue FDR control over the overall batch of discoveries. This decentralized testing of different sequences using different algorithms is particularly aligned with the online FDR setup, where not all hypotheses are known in advance. In fact, this "online" property of the mFDR was recognized even within offline FDR control (van den Oord, 2008).

To simplify our presentation, we will often suppress the distinction, referring to both of these metrics as "FDR."

In online FDR control, the set of rejections possibly changes at each time step, implying changes in mFDR and FDR. Therefore, in online settings, we have to consider $\mathcal{R}(t)$, which is the set of rejections up to time t, and the naturally implied $\text{mFDR}(t)$ and $\text{FDR}(t)$. We will also use the symbol $\mathcal{V}(t) := \mathcal{R}(t) \cap \mathcal{H}^0$ to denote the set of false rejections made up to time t. The main objective of online FDR algorithms is to ensure $\text{mFDR}(t) \leq \alpha$ or $\text{FDR}(t) \leq \alpha$, for a chosen level α and *for all times* t.

Many of the online FDR algorithms that have been proposed to date in the literature are special cases of the generalized alpha-investing (GAI) framework (Aharoni and Rosset, 2014). The initial interpretation of these algorithms, as put forward by Foster and Stine (2008), relied on a notion of dynamically changing "alpha-wealth." Ramdas et al. (2017, 2018) subsequently presented an alternative perspective on GAI algorithms. In this view, GAI algorithms are viewed as keeping track of an empirical estimate of the true false discovery proportion, denoted $\widehat{\text{FDP}}(t)$, and they assign test levels α_t in a way that ensures $\widehat{\text{FDP}}(t) \leq \alpha$ for all time steps t, where α is the pre-specified FDR level. In the earlier paper (Ramdas et al., 2017), they show that such control of FDP estimates also yields FDR control. This perspective—which is equivalent to the earlier, wealth characterization of GAI algorithms—will provide the mathematical framework upon which we build in this paper.

Finally, we recap the typical assumptions made for null p-values in the FDR literature. If a hypothesis H_i is truly null, then the corresponding p-value P_i is stochastically larger than the uniform distribution ("super-uniformly distributed," or "super-uniform" for short), meaning that:

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

This condition is sometimes generalized to the online FDR setting by incorporating a filtration $\mathcal{F}^{i-1}$, resulting in the following assumption:

$$\text{If the null hypothesis } H_i \text{ is true, then } \Pr\big\{P_i \leq u \;\big|\; \mathcal{F}^{i-1}\big\} \leq u \text{ for all } u \in [0,1], \tag{2}$$

Here, $\mathcal{F}^i$ captures all relevant information about the first i tests. As we discuss in later sections, however, this condition can be overly stringent when there are interactions between p-values, and we will accordingly introduce weaker super-uniformity assumptions.

1.2 Problem formulation and contribution

We now give a formal introduction to the problem setting, at the same time introducing the necessary notation for the sections to follow.

At time step $t \in \mathbb{N}$, the test of hypothesis H_t begins, and the p-value resulting from this test is denoted P_t. In contradistinction to the standard online FDR paradigm, P_t is not required to be known at time t; indeed, this test is not fully *performed* at time t, but is only initiated at time t. The *decision time* for H_t is denoted E_t; this is the time of possible rejection. Fully synchronous testing is thus an instance of this setting in which $E_t = t$, as assumed in classical online FDR work. In general, however, $E_t \neq t$. Note also that, unlike in the classical online FDR problem, the set of rejections $\mathcal{R}(t)$ and false rejections $\mathcal{V}(t)$ at time t now consider not all $\{P_i : i \leq t\}$, but only $\{P_i : E_i \leq t\}$:

$$\mathcal{R}(t) = \{i \in [t] : E_i \leq t, H_i \text{ is rejected}\},\ \mathcal{V}(t) = \mathcal{R}(t) \cap \mathcal{H}^0.$$

In addition, to capture the desideratum of statistical validity in the face of data reuse, we allow the possibility of the p-values not being completely independent; in particular, we allow *local dependence*. Here, we envision P_t having arbitrary, possibly adversarial dependence on $P_{t-1}, \ldots, P_{t-L_t}$, while the dependence between

P_t and P_j, $j < t - L_t$ is limited. For simplicity the reader can assume $P_t \perp P_{t-L_t-1}, P_{t-L_t-2}, \ldots, P_1$, however in later sections we will discuss some restricted forms of dependence between P_t and P_j, for $j < t - L_t$, handled by our results.

We treat E_t as fixed but unknown before time E_t itself. While the p-value and the duration of a test could indeed be dependent random quantities—for example, when the duration is a reasonably good proxy for sample size—here E_t is *not* the absolute duration on a meaningful time scale, but it merely captures how many tests have started before the decision for the t-th test. Thus, treating E_t as fixed roughly corresponds to asserting independence between P_t and the number of newly created tests before test t finishes. As we envision a highly decentralized setting with little between-test coordination, we deem this assumption reasonable. We do, however, acknowledge the possibility of a more coordinated setting with P_t and E_t randomly coupled, and this is an important avenue for future work.

Under the setup described above, the goal is to produce test levels α_t dynamically at the beginning of the t-th test, such that, *despite arbitrary local dependence and regardless of the decision times* E_t, the false discovery rate is controlled at any given moment under a pre-specified level α. In this work, we provide procedures which achieve this goal, both at all fixed times $t \in \mathbb{N}$, as well as adaptively chosen stopping times.

It is important to remark that, even under independence of p-values, we cannot simply ignore the asynchronous aspects of the problem and naively apply an existing online FDR algorithm. We discuss two such plausible but naive applications of online methodology, and discuss why they are invalid.

One natural adjustment could be to apply an online FDR algorithm whenever each test finishes (that is, whichever test is the t-th one to finish, test it at level α_t). This scheme would only assign α_t to a test at the end of that test, which is unrealistic because sequential hypothesis tests—parametric tests such as Wald's sequential probability ratio test (SPRT), and nonparametric tests as well (Balsubramani and Ramdas, 2016)—typically require specification of the target type I error level in advance because it is an important component of their stopping rule. For examples of sequential tests in clinical trials, see, e.g., (Bartroff et al., 2012; Bartroff and Lai, 2008; Bartroff and Song, 2014). The same constraint holds for more recent, multi-armed bandit approaches to A/B testing (Yang et al., 2017; Jamieson and Jain, 2018). Thus, we need to specify α_t at the *start* of test t. That said, there do exist tests which only require the test level at decision time. If *all* tests in the sequence are of the latter kind, existing online methodology is indeed sufficient; however, we find this assumption too strong, especially given the popularity of bandit approaches in modern testing applications.

Alternatively, one could imagine computing α_t at the start of test t by applying an online FDR algorithm to the completed tests only, and ignoring those that have not finished. From the theory perspective, this clearly comes with no formal guarantee; for example, one could imagine starting N tests, and all N tests finishing at once, after the N-th test has started. Given that there are no completed tests at the time of test level assignment, all tests would receive the same level, which would violate the FDR requirement under any natural setting (we ignore trivial special cases such as all hypotheses being non-null). Somewhat less trivially, Figure 3 plots the FDR and mFDR achieved by this naive heuristic in a simulation setting from Section 7. When the proportion of non-null p-values is relatively small—as one would generally expect in practice—this heuristic severely violates the FDR requirement.

1.3 Related work

There is a vast literature on sequential testing (see, e.g., Wald, 1945; Chernoff, 1959; Albert, 1961; Naghshvar and Javidi, 2013). We do not aim to contribute to that literature per se; rather, our goal is to consider multiple testing through a more realistic lens as an outer sequential process, one that acknowledges the existence of inner sequential processes that are based on sequential testing.

Likewise, there is a large and growing literature on false discoveries in multiple testing, aimed at solving a range of problems, often addressing issues of scientific reproducibility in research (Ioannidis, 2005). Here we focus on work whose methods or objectives have the most overlap with ours. In particular, we focus on literature on "online" methods in multiple testing, and compare and contrast those solutions to the ones we propose.

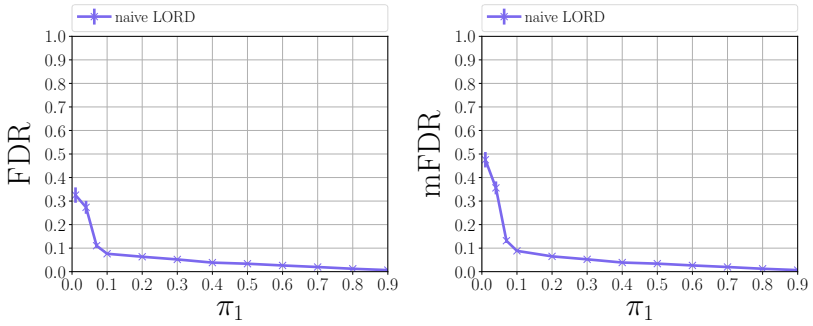

Figure 3: FDR and mFDR achieved by a naive application of the LORD algorithm with target FDR level $\alpha = 0.05$ (Javanmard and Montanari, 2018) in an asynchronous environment. We adopt the experimental setting from Section 7; we set the asynchrony parameter to $p = 1/150$, and the mean of observations under the alternative to $\mu_c = 3$. Here, π_1 is the proportion of tested hypotheses which are non-null. Both FDR and mFDR are controlled only for $\pi_1 \geq 0.4$.

The most salient difference is that we address the general problem of asynchrony; when there is no asynchrony, meaning $E_t = t$, our approach recovers a slew of existing methods, including work by Foster and Stine (2008), Aharoni and Rosset (2014), Javanmard and Montanari (2015, 2018), Ramdas et al. (2017, 2018).

Most previous work also differs from ours in that it assumes that condition (2) holds. This condition is too strong for the notion of local dependence this paper considers; indeed, in Section 4 we present a simple toy example in which this assumption fails. An exception is the work of Javanmard and Montanari (2015, 2018), who discuss sufficient conditions for achieving FDR control under arbitrary dependence within the p-value sequence. However, these conditions essentially imply an alpha-spending-like correction for the test levels, making their proposed procedure overly conservative. We elaborate on this argument and empirically demonstrate this observation in Section 7.

Robertson and Wason (2018) have investigated the performance of several online FDR algorithms empirically, including all of those listed above, when the p-value sequence is positively dependent. They do not, however, provide any formal guarantees for those procedures that have thus far been shown to work only under independence. We make partial progress to justifying their empirical observations by proving that LOND provably controls FDR under positive dependence.

Recently, there has also been some work specifically motivated by controlling false discoveries in A/B testing in the tech industry (Yang et al., 2017). However, their setup was again fully synchronous, and assume that the observations are independent across all experiments, which are the two assumptions this paper deems too strong and circumvents.

The vast literature on adaptive data analysis (Dwork et al., 2015b,a; Bassily et al., 2016; Blum and Hardt, 2015) focuses on an online setting where a distribution is adaptively queried for a chosen functional, and at each step these queries are answered by making use of a single data set coming from that distribution. This line of work also has the goal of preventing false discovery, however by proving generalization bounds, rather than controlling the FDR in online multiple testing.

Ordered hypothesis testing considers tests for which additional prior information is available, and allows sorting null hypotheses from least to most promising (Li and Barber, 2017; Lei and Fithian, 2016; Lynch et al., 2017; G'Sell et al., 2016). In these papers, however, the word "sequential" or "ordered" does not refer to online testing; these methods are set in an offline environment, requiring access to all p-values at once. In our approach, we allow testing a possibly infinite number of hypotheses with no available knowledge of the future p-values.

2. Conflict sets: the unifying approach

In this section we describe a general, abstract formulation of multiple testing under asynchrony and dependence, which unifies the seemingly disparate solutions of this paper and provides the point of departure for deriving specific algorithms. We describe two such procedures, which we will refer to as LORD* and SAFFRON*, that control mFDR within this framework.

LORD* and SAFFRON* build off the LORD (Javanmard and Montanari, 2018) and SAFFRON (Ramdas et al., 2018) algorithms. Like SAFFRON, SAFFRON* allows the user to choose a parameter $\lambda_t \geq \alpha_t$, which is the "candidacy threshold" at time t, meaning that, if $P_t \leq \lambda_t$, then P_t is referred to as a *candidate* for rejection. We will discuss this extension introduced in the SAFFRON procedure further below; for now, we simply note that it is an analog of the notion of "null-proportion adaptivity" in the offline multiple testing literature. Indeed, Ramdas et al. (2018) argue that LORD can be seen as the online analog of the BH procedure (Benjamini and Hochberg, 1995), while SAFFRON can be seen as the online analog of the adaptive Storey-BH procedure (Storey, 2002; Storey et al., 2004).

Throughout we let $R_t := \mathbf{1}\{P_t \leq \alpha_t\}$ denote the *indicator for rejection*, and $C_t := \mathbf{1}\{P_t \leq \lambda_t\}$ denote the *indicator for candidacy*.

We now define several filtrations, which capture the increasing information available to the experimenter as well as the FDR algorithm.

By $\mathcal{L}^t$, we denote a filtration that captures all relevant information about the tests that started up to, and including, time t, for the LORD* procedure. Formally, $\mathcal{L}^t := \sigma(\{R_1, \ldots, R_t, \})$. For SAFFRON*, we also incorporate candidates in the filtration: $\mathcal{S}^t := \sigma(\{R_1, C_1, \ldots, R_t, C_t\})$. Many of our arguments will apply to both algorithms; we accordingly use $\mathcal{F}^t$ to indicate a generic filtration that can be either $\mathcal{L}^t$ or $\mathcal{S}^t$.

With each test and its corresponding hypothesis, we associate a *conflict set*. For the test starting at step t, we denote this set $\mathcal{X}^t$; it consists of a (not necessarily strict) subset of $\{1, \ldots, t-1\}$. For example, $\mathcal{X}^5$ could be $\{3, 4\}$. The reason why we refer to this set as *conflicting* for test t is because it contains the indices of tests that interact with the t-th test in some unknown way. This could mean that, at time t, there is missing information about these tests, or that there potentially exists some arbitrary dependence between those tests and the upcoming one. More explicitly, we let

$$\mathcal{X}^t = \{i \in [t-1] \ : \ E_i \geq t\} \cup \{t - L_t, \ldots, t-1\},$$

where L_t is the sequence of dependence lags. In words, $\mathcal{X}^t$ consists of all tests that have not finished running or are locally dependent with test t.

We require the conflict sets to be *monotone*: each index t has to be in a continuous "block" of conflict sets. More formally, if there exists j such that $t \in \mathcal{X}^j$, then $t \in \mathcal{X}^i$, for all $i \in \{t+1, \ldots, j\}$. Without any constraint on the sequence $\{L_t\}$, the conflict sets need not be monotone. Therefore, we translate the condition of monotonicity of conflict sets into a constraint on the sequence $\{L_t\}$ as: $L_{t+1} \leq L_t + 1$. Informally, this is just a requirement that the "non-conflicting information" does not decrease with time. This will ensure that the test level α_t and candidacy threshold λ_t have at least as much knowledge about prior tests as α_{t-1} and λ_{t-1}. Moreover, this requirement is indeed a natural one, and usual testing practices satisfy it; for example, this condition holds if dependent p-values come in disjoint blocks.

We define the *last-conflict time* of test t as $\tau_t := \max\{j : t \in \mathcal{X}^j\}$. If test t never appears in a conflict set, we take $\tau_t = t$.

Consider again the filtration $\mathcal{F}^t$. A subtlety we initially ignored is that the superscript t does not correspond to the physical quantity of time. In particular, different tests may run for different lengths of time and the decision time for each test may even be random; therefore, R_t might be known before R_{t-1}. This motivates us to define a filtration as a counterpart of $\mathcal{F}^t$ whose increase at each step corresponds to the real increase in knowledge with time. We introduce $\mathcal{F}^{-\mathcal{X}^t}$ as the *non-conflicting filtration*; the sigma-algebra $\mathcal{F}^{-\mathcal{X}^t}$ contains information about the tests that started before time t which are *not* in the conflict set of test t. In particular, $\mathcal{L}^{-\mathcal{X}^t} := \sigma(\{R_i : i \leq t-1, i \notin \mathcal{X}^t\})$ for LORD*, $\mathcal{S}^{-\mathcal{X}^t} := \sigma(\{R_i, C_i : i \leq t-1, i \notin \mathcal{X}^t\})$ for SAFFRON*, and again we use $\mathcal{F}^{-\mathcal{X}^t}$ to generically denote either $\mathcal{L}^{-\mathcal{X}^t}$ or $\mathcal{S}^{-\mathcal{X}^t}$. We have that $\mathcal{F}^{-\mathcal{X}^t} \subseteq \mathcal{F}^{t-1}$. Notice that we promised to make this set a *filtration*; if $\mathcal{X}^t$ was an arbitrary set of indices, this would not in

general be satisfied. However, it is straightforward to verify that the monotonicity property of conflict sets ensures that $\mathcal{F}^{-\mathcal{X}^t}$ indeed forms a filtration.

We will design α_t and λ_t to be $\mathcal{F}^{-\mathcal{X}^t}$-measurable. This is essentially the idea of pessimism mentioned earlier—among all tests that finished before the t-th one starts, α_t and λ_t have to ignore the ones conflicting with test t in order to guard against unknown interactions that the conflicting tests have with the upcoming one.

Finally, we will generally require the following super-uniformity condition for null p-values:

$$\text{If the null hypothesis } H_t \text{ is true, then } \Pr\Big\{P_t \le u \;\Big|\; \mathcal{F}^{-\mathcal{X}^{E_t}}\Big\} \le u, \text{ for all } u \in [0,1]. \tag{3}$$

This is a condition that requires validity of null p-values: given the knowledge one has before making a decision, if a hypothesis is truly null, it has to be well-behaved. However, unlike in classical online FDR work, we do not have $\mathcal{F}^{-\mathcal{X}^{E_t}} = \mathcal{F}^{t-1}$. As we discuss further in later sections, assumption (3) will allow arbitrary local dependence, as well some limited, but nevertheless important, forms of dependence between distant p-values. Note that, if the distant p-values are independent—a setting we study in Section 4 and Section 5—this condition is automatically satisfied.

2.1 The LORD* algorithm

Following a recently proposed framework (Ramdas et al., 2017), we define LORD* and SAFFRON* as arbitrary update rules which control a certain estimate of the false discovery proportion under a pre-specified level α; the two algorithms differ in their choice of estimate. In Subsection 2.3, we introduce additional analysis tools which will justify the choice of these estimates.

LORD* is defined as any update rule for α_t that ensures that the estimate

$$\widehat{\text{FDP}}_{\text{LORD*}}(t) := \frac{\sum_{j\le t}\alpha_j}{(\sum_{j\le t, j\notin\mathcal{X}^t} R_j)\vee 1}.$$

is at most α for all $t \in \mathbb{N}$.

Below we state two different versions of LORD*, using two different test level updates. Algorithm 1 generalizes the LORD++ procedure (Javanmard and Montanari, 2018; Ramdas et al., 2017), while Algorithm 2 generalizes its predecessor, the LOND procedure (Javanmard and Montanari, 2015). These are not the only ways of assigning α_j that are consistent with the assumptions and satisfy the definition of LORD* in their control of $\widehat{\text{FDP}}_{\text{LORD*}}$, but they are our focus in the remainder of the paper. Other rules can be developed as extensions of the rules in the LORD paper (Javanmard and Montanari, 2018).

To state the algorithms in this paper, we will make use of the variable r_k, which refers to the first time that k rejections are non-conflicting, meaning that there exist k rejected hypotheses which are no longer in the conflict set at that time. That is, we define r_k as:[2]

$$r_k := \min\{i \in [t] : \sum_{j=1}^{i} R_j \mathbf{1}\{\tau_j \le i\} \ge k\}. \tag{4}$$

Algorithm 1 The LORD++ algorithm under general conflict sets (a special case of LORD*)

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \le \alpha$
Set $\alpha_1 = \gamma_1 W_0$
for $t = 1, 2, \ldots$ **do**
  start t-th test with level α_t
  $\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-r_1}(\alpha - W_0) + \left(\sum_{j\ge 2}\gamma_{t+1-r_j}\right)\alpha$
end

2. Here, as well as in the rest of this paper, we define the minimum of an empty set to be $-\infty$.

Algorithm 2 The LOND algorithm under general conflict sets (a special case of LORD*)

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$
Set $\alpha_1 = \gamma_1 \alpha$
for $t = 1, 2, \ldots$ **do**
  start t-th test with level α_t
  $\alpha_{t+1} = \alpha \gamma_{t+1} \left((\sum_{j=1}^{t} \mathbf{1}\{P_j \le \alpha_j, \tau_j \le t\}) \vee 1 \right)$
end

It is a simple algebraic exercise to verify that the two update rules given for α_t indeed guarantee that $\widehat{\text{FDP}}_{\text{LORD*}}(t) \le \alpha$ for all $t \in \mathbb{N}$.

2.2 The SAFFRON* algorithm

In response to LORD and LOND's FDP estimate, the SAFFRON method was derived after observing that the former might be overly conservative estimates of the FDP. Indeed, if the tested sequence contains a significant fraction of non-nulls, and if the non-nulls yield strong signals for rejection, the realized FDP and the estimated FDP might be very far apart. Motivated by this observation, SAFFRON was developed as the adaptive counterpart of LORD which keeps track of an empirical estimate of the null proportion, similar to the way in which Storey et al. (Storey, 2002; Storey et al., 2004) improved upon the BH procedure (Benjamini and Hochberg, 1995). We thus propose the SAFFRON* algorithm to maintain control over the following estimate:

$$\widehat{\text{FDP}}_{\text{SAFFRON*}}(t) := \frac{\sum_{j<t, j \notin \mathcal{X}^t} \frac{\alpha_j}{1-\lambda_j} \mathbf{1}\{P_j > \lambda_j\} + \sum_{j \in \{\mathcal{X}^t \cup \{t\}\}} \frac{\alpha_j}{1-\lambda_j}}{(\sum_{j \le t, j \notin \mathcal{X}^t} R_j) \vee 1}.$$

Any update rule for α_t and λ_t ensuring $\widehat{\text{FDP}}_{\text{SAFFRON*}}(t) \le \alpha$ for all $t \in \mathbb{N}$ satisfies the definition of SAFFRON*. Algorithm 3 and Algorithm 4 describe two particular instances of SAFFRON*, obtained for specific choices of the sequence $\{\lambda_j\}$. We present an algorithmic specification of SAFFRON* for the constant sequence $\{\lambda_j\} \equiv \lambda$ in Algorithm 3. A different case of SAFFRON* is presented in Algorithm 4, where we use the alpha-investing strategy $\lambda_j = \alpha_j$ (Foster and Stine, 2008; Ramdas et al., 2018).

For the updates below, recall the definition of r_k from equation (4).

Algorithm 3 The SAFFRON* algorithm for constant λ under general conflict sets

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, candidate threshold $\lambda \in (0, 1)$, initial wealth $W_0 \le \alpha$
$\alpha_1 = (1-\lambda)\gamma_1 W_0$
for $t = 1, 2, \ldots$ **do**
  start t-th test with level α_t
  $\alpha_{t+1} = \min\left\{\lambda, (1-\lambda)\left(W_0 \gamma_{t+1-C_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-C_{1+}} + \sum_{j \ge 2} \alpha \gamma_{t+1-r_j-C_{j+}}\right)\right\}$,
  where $C_{j+} = \sum_{i=r_j+1}^{t} C_i \mathbf{1}\{i \notin \mathcal{X}^t\}$
end

Algorithm 4 The alpha-investing algorithm under general conflict sets (a special case of SAFFRON*)

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \le \alpha$
$s_1 = \gamma_1 W_0$
$\alpha_1 = s_1/(1+s_1)$
for $t = 1, 2, \ldots$ **do**
  start t-th test with level α_t
  $s_{t+1} = W_0 \gamma_{t+1-R_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-R_{1+}} + \sum_{j \ge 2} \alpha \gamma_{t+1-r_j-R_{j+}}$, where $R_{j+} = \sum_{i=r_j+1}^{t} R_i \mathbf{1}\{i \notin \mathcal{X}^t\}$
  $\alpha_{t+1} = s_{t+1}/(1+s_{t+1})$,
end

2.3 Oracle estimate under conflict sets

Following Ramdas et al. (2018), we analyze LORD* and SAFFRON* through an *oracle estimate* of the false discovery proportion. This quantity serves as a good estimate of the true false discovery proportion, and controlling it under a pre-specified level guarantees that FDR is also controlled. Let the oracle estimate of the FDP be defined as:

$$\text{FDP}^*(t) := \frac{\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j}{(\sum_{E_j \le t} R_j) \vee 1},$$

where we recall that α_j is required to be $\mathcal{F}^{-\mathcal{X}^j}$-measurable, across all j. The following proposition gives formal justification for using $\text{FDP}^*(t)$ as a proxy for the true FDP.

Proposition 1 *Suppose that the null p-values are super-uniform conditional on $\mathcal{F}^{-\mathcal{X}^{E_t}}$, meaning $Pr\Big\{P_t \le u \;\Big|\; \mathcal{F}^{-\mathcal{X}^{E_t}}\Big\} \le u$, for all $u \in [0,1]$ and $t \in \mathcal{H}^0$. Then, for all times $t \in \mathbb{N}$, the condition $\text{FDP}^*(t) \le \alpha$ implies that $\text{mFDR}(t) \le \alpha$.*

Note that Proposition 1 is technically true even if we only consider tests for which $E_j \le t, j \in \mathcal{H}^0$ in the numerator of $\text{FDP}^*(t)$. However, it is not clear how to achieve this without ensuring $\text{FDP}^*(t) \le \alpha$. For example, even if $\frac{\sum_{E_j \le t, j \in \mathcal{H}^0} \alpha_j}{(\sum_{E_j \le t} R_j) \vee 1} \le \alpha$ at time t, it is possible that in subsequent rounds all tests will finish without any new rejections, thus increasing the FDP estimate. Therefore, we need to assign α_j conservatively, such that this estimate is provably controlled under α, despite unknown future outcomes which might augment the FDP estimate.

The fact that α_t is measurable with respect to $\mathcal{F}^{-\mathcal{X}^t}$ should give us pause. Even though we are only required to guarantee $\text{FDP}^*(t) \le \alpha$, we cannot rely on the rejection indicators that push down the value of $\text{FDP}^*(t)$, if they are in the current conflict set. As a consequence, α_t has to ensure $\text{FDP}^*(t) \le \alpha$ for the *worst-case* configuration of conflicting rejections; that is, when $R_j = 0$ for all $j \in \mathcal{X}^t$. This motivates us to define the oracle estimate of the FDP *under conflict sets*:

$$\text{FDP}^*_{\text{conf}}(t) := \frac{\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j}{(\sum_{j \le t, j \notin \mathcal{X}^t} R_j) \vee 1}. \tag{5}$$

Since this quantity is only more conservative than the oracle estimate, controlling it under α will preserve the guarantees given by Proposition 1. However, notice an unfortunate fact about both oracle estimates—they depend on the unobservable set $\mathcal{H}^0$. This implies that not even $\text{FDP}^*_{\text{conf}}(t)$ can be controlled tightly. For this reason, LORD* and SAFFRON* construct *empirical* estimates of $\text{FDP}^*_{\text{conf}}(t)$, such that the properties given in Proposition 1 are retained. For LORD*, claiming mFDR control at fixed times boils down to a simple observation: for any chosen α, $\text{FDP}^*_{\text{conf}}(t) \le \widehat{\text{FDP}}_{\text{LORD*}}(t) \le \alpha$, hence by Proposition 1 mFDR is controlled. SAFFRON* controls mFDR by virtue of ensuring that, on average, $\text{FDP}^*_{\text{conf}}(t) \le \alpha$. We make this argument formal in Section 6.

3. Example 1: Asynchronous online FDR control

In this section, we look at one instantiation of the conflict-set framework, which considers arbitrary asynchrony but limits possible dependencies between p-values. This immediately gives two procedures for asynchronous online testing as special cases of LORD* and SAFFRON*. From here forward we will refer to these methods as $\text{LORD}_{\text{async}}$ and $\text{SAFFRON}_{\text{async}}$, respectively. In Section 6, we provide mFDR guarantees of these procedures in terms of the general conflict-set setting, as well as additional FDR guarantees for $\text{LORD}_{\text{async}}$ and $\text{SAFFRON}_{\text{async}}$ under a strict independence assumption.

In this section, the only conflicting tests are those whose outcomes are unknown, since the allowed dependencies will be fairly restrictive. Therefore, the asynchronous conflict set at time t is:

$$\mathcal{X}^t_{\text{async}} = \{i \in [t-1] : E_i \ge t\},$$

which is observable at time $t-1$. This simplified conflict set implies that the last-conflict time of test t is, naturally, $\tau_t = E_t$.

Denote by $\mathcal{R}_t$ the *set of rejections* at time t, and similarly let $\mathcal{C}_t$ denote the *set of candidates* at time t:

$$\mathcal{R}_t = \{i \in [t] : E_i = t, P_i \leq \alpha_i\},\ \mathcal{C}_t = \{i \in [t] : E_i = t, P_i \leq \lambda_i\}.$$

Therefore, $\mathcal{R}(t) = \cup_{i=1}^{t} \mathcal{R}_t$. With this, we can write the non-conflicting filtrations $\mathcal{L}_{\text{async}}^{-\mathcal{X}^t}$ and $\mathcal{S}_{\text{async}}^{-\mathcal{X}^t}$ compactly as:

$$\mathcal{L}_{\text{async}}^{-\mathcal{X}^t} := \sigma(\mathcal{R}_1, \ldots, \mathcal{R}_{t-1}),\ \mathcal{S}_{\text{async}}^{-\mathcal{X}^t} := \sigma(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{t-1}, \mathcal{C}_{t-1}).$$

Since the arguments for LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ have significant overlap, for brevity we write $\mathcal{F}_{\text{async}}^{-\mathcal{X}^t}$ to refer to both $\mathcal{L}_{\text{async}}^{-\mathcal{X}^t}$ and $\mathcal{S}_{\text{async}}^{-\mathcal{X}^t}$, where possible. Recall from Section 2 that α_t is designed to be measurable with respect to $\mathcal{F}_{\text{async}}^{-\mathcal{X}^t}$; here this essentially means that it is computed as a function of the outcomes known by time t. For SAFFRON$_{\text{async}}$, additionally λ_t is $\mathcal{S}_{\text{async}}^{-\mathcal{X}^t}$-measurable. More generally, for LORD$_{\text{async}}$, we can choose $\alpha_t = f_t(\mathcal{R}_1, \ldots, \mathcal{R}_{t-1})$, for any deterministic function f_t as long as the correct FDP estimate is controlled. The SAFFRON$_{\text{async}}$ procedure also keeps track of encountered candidates, hence we can take $\alpha_t = g_t(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{t-1}, \mathcal{C}_{t-1})$ and $\lambda_t = h_t(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{t-1}, \mathcal{C}_{t-1})$, for deterministic functions g_t and h_t.

Our mFDR guarantees hold under a condition which we term *asynchronous super-uniformity*:

$$\text{If the null hypothesis } H_t \text{ is true, then } \Pr\left\{ P_t \leq u \;\middle|\; \mathcal{F}_{\text{async}}^{-\mathcal{X}^{E_t}} \right\} \leq u, \text{ for all } u \in [0,1]. \tag{6}$$

This condition essentially shapes the allowed dependencies between p-values. It is immediately implied if the p-values are independent. However, it is strictly weaker. For example, it allows revisiting p-values which were previously not rejected. Suppose we have tested independent p-values thus far, and we failed to reject H_t, that is $P_t > \alpha_t$. If at a later time $s > t$ we have a higher error budget $\alpha_s > \alpha_t$, we can, somewhat surprisingly, test H_t using the *same* p-value P_t again at time s. This clearly violates independence of P_t and P_s (as they are identical), however condition (6) is nevertheless satisfied. Indeed, for all $u > \alpha_t$:

$$\Pr\left\{ P_s \leq u \;\middle|\; \mathcal{F}_{\text{async}}^{-\mathcal{X}^{E_s}} \right\} = \Pr\{P_s \leq u \mid P_s > \alpha_t\} \leq \frac{u - \alpha_t}{1 - \alpha_t} \leq u,$$

where the equality follows because $P_s \perp \mathcal{F}_{\text{async}}^{-\mathcal{X}^{E_s}} \mid \mathbf{1}\{P_s > \alpha_t\}$. On the other hand, if $u \leq \alpha_t$, $\Pr\{P_s \leq u \mid P_s > \alpha_t\} = 0 \leq u$, and hence condition (6) follows.

The LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ algorithms

We turn to an analysis of how the abstract LORD* and SAFFRON* procedures translate into our asynchronous testing scenario, for the particular choice of conflict set $\mathcal{X}_{\text{async}}^t$. They utilize all available information; the conflict set—the tests whose outcomes the algorithms ignore—consists only of the tests about which we temporarily lack information.

Plugging in the definition of $\mathcal{X}_{\text{async}}^t$, we obtain the following empirical estimate of the false discovery proportion for LORD$_{\text{async}}$:

$$\widehat{\text{FDP}}_{\text{LORD}_{\text{async}}}(t) = \frac{\sum_{j \leq t} \alpha_j}{(\sum_{j \leq t} \mathbf{1}\{P_j \leq \alpha_j, E_j < t\}) \vee 1}.$$

For SAFFRON$_{\text{async}}$, we obtain the following estimate:

$$\widehat{\text{FDP}}_{\text{SAFFRON}_{\text{async}}}(t) = \frac{\sum_{j \leq t} \frac{\alpha_j}{1-\lambda_j} (\mathbf{1}\{P_j > \lambda_j, E_j < t\} + \mathbf{1}\{E_j \geq t\})}{(\sum_{j \leq t} \mathbf{1}\{P_j \leq \alpha_j, E_j < t\}) \vee 1}.$$

Consider the dynamics of these two algorithms, and how their pessimism comes into play. Whenever a test starts, they increase their FDP estimate, expecting that the resulting p-value will have no favorable contribution. However, when the test in question ends, they readjust the FDP estimate if they see a positive outcome, namely a candidate and/or rejection. This shows that testing in parallel indeed has a cost—due to pessimistic expectations about the tests in progress, the algorithms remain conservative when assigning a new test level. For this reason, asynchronous testing should be used with caution, and the number of tests run in parallel should be monitored closely. Indeed, in the asymptotic limit where the number of parallel tests tends to infinity, the algorithm behaves like alpha-spending; i.e., the sum of all assigned test levels converges to the error budget α.

Substituting $\mathcal{X}^t$ for $\mathcal{X}^t_{\text{async}}$ in Algorithms 1-4 yields procedures for asynchronous online FDR control. The explicit statements of these algorithms, which correspond to asynchronous versions of LORD++, LOND, SAFFRON, and alpha-investing, are given in the Appendix.

4. Example 2: Online FDR control under local dependence

In this section, we derive online FDR procedures that handle local dependencies. We begin with the fully synchronous setting studied in classical online FDR literature, and turn to the asynchronous environment in the next section.

A standard assumption in existing work on online FDR has been independence of p-values, a requirement that is rarely justified in practice. Tests that cluster in time often use the same data, null hypotheses depend on the outcomes of recent tests, etc. On the other hand, arbitrary dependence between any two p-values in the sequence is also arguably unreasonable—very old data used for testing in the past is usually considered "stale," and hypotheses tested a long time ago may bear little relevance to current hypotheses. In light of this, we consider a notion of *local dependence*:

$$\text{for all } t > 0, \text{ there exists } L_t \in \mathbb{N} \text{ such that } P_t \perp P_{t-L_t-1}, P_{t-L_t-2}, \ldots, P_1,$$

where $\{L_t\}$ is a fixed sequence of parameters which we refer to as lags.

Since we allow P_t to have arbitrary dependence on the previous L_t p-values, some of these dependencies might be adversarial toward the statistician, and, with "peeking" into this adversarial set, the nulls might no longer behave super-uniformly. Suppose we observe a sample $X \sim N(\mu, 1)$, and wish to test two hypotheses using this sample. Let the two hypotheses be $H_1 : \mu < 0$ and $H_2 : \mu \geq 0$. If, for instance, $R_1 = 0$, we know that $P_2 \leq 1 - \alpha_1$ almost surely, implying that P_2 is not super-uniform, given the information about past tests. On the other hand, if we were to ignore the outcome of the first test, P_2 would indeed be super-uniform.

This observation motivates us to define the conflict set for testing under local dependence as:

$$\mathcal{X}^t_{\text{dep}} := \{t - L_t, \ldots, t - 1\}.$$

The non-conflicting filtrations $\mathcal{L}^{-\mathcal{X}^t}_{\text{dep}}$ for LORD$_{\text{dep}}$ and $\mathcal{S}^{-\mathcal{X}^t}_{\text{dep}}$ for SAFFRON$_{\text{dep}}$ are respectively given by:

$$\mathcal{L}^{-\mathcal{X}^t}_{\text{dep}} := \sigma(R_1, \ldots, R_{t-L_t-1}),\ \mathcal{S}^{-\mathcal{X}^t}_{\text{dep}} := \sigma(R_1, C_1, \ldots, R_{t-L_t-1}, C_{t-L_t-1}).$$

Since most formal arguments in this section apply to both procedures, we use $\mathcal{F}^{-\mathcal{X}^t}_{\text{dep}}$ to indicate that the filtration in question could be both $\mathcal{L}^{-\mathcal{X}^t}_{\text{dep}}$ and $\mathcal{S}^{-\mathcal{X}^t}_{\text{dep}}$.

In contrast to asynchronous testing, the levels α_t and λ_t under local dependence ignore some portion of available information, specifically the outcomes of the last L_t tests. Notice the difference between these two settings—in the asynchronous setting, pessimism guards against *unknown outcomes*, while here pessimism guards against *known outcomes*. Perhaps counterintuitively, this observation means that the pessimism of LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ actually guards against possible disadvantageous direct impact of the last L_t p-values on the upcoming one. In the Appendix we instantiate the test levels and candidacy thresholds according to Algorithms 1-4, however more generally we allow $\alpha_t = f_t(R_1, \ldots, R_{t-L_t-1})$ for LORD$_{\text{dep}}$, and $\alpha_t = g_t(R_1, C_1, \ldots, R_{t-L_t-1}, C_{t-L_t-1})$ and $\lambda_t = h_t(R_1, C_1, \ldots, R_{t-L_t-1}, C_{t-L_t-1})$ for SAFFRON$_{\text{dep}}$.

Consider some P_t which is from a null hypothesis. As previously emphasized, we cannot trust P_t to behave like a true null, given that we already know its last L_t predecessors that have a direct impact on it. The appropriate super-uniformity condition satisfied by locally dependent p-values thus ignores these last L_t p-values and is of the following form:

$$\text{If the null hypothesis } H_t \text{ is true, then } \Pr\Big\{ P_t \le u \;\Big|\; \mathcal{F}_{\text{dep}}^{-\mathcal{X}^t} \Big\} \le u, \text{ for all } u \in [0,1]. \tag{7}$$

This will allow setting $\alpha_t \in \mathcal{F}_{\text{dep}}^{-\mathcal{X}^t}$, while knowing $\Pr\Big\{ P_t \le \alpha_t \;\Big|\; \mathcal{F}_{\text{dep}}^{-\mathcal{X}^t} \Big\} \le \alpha_t$. Importantly, unlike in the previous section where the appropriate super-uniformity condition implied dependence constraints on the p-values, condition (7) is immediately true by local dependence.

The LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ algorithms

As in Section 3, we analyze the particular instances of LORD* and SAFFRON* that are obtained by taking the conflict set of Section 2 to be $\mathcal{X}_{\text{dep}}^t = \{t - L_t, \dots, t-1\}$. Since this conflict set is deterministic, unlike $\mathcal{X}_{\text{async}}^t$, the estimate of the false discovery proportion that LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ keep track of is completely determined L_t steps ahead, that is at time $t - L_t - 1$.

By definition of the general estimates and the conflict set in consideration, LORD$_{\text{dep}}$ controls the following quantity:

$$\widehat{\text{FDP}}_{\text{LORD}_{\text{dep}}}(t) = \frac{\sum_{j \le t} \alpha_j}{(\sum_{j \le t, j \notin \{t-L_t,\dots,t-1\}} R_j) \vee 1}.$$

The SAFFRON$_{\text{dep}}$ method, on the other hand, controls the estimate:

$$\widehat{\text{FDP}}_{\text{SAFFRON}_{\text{dep}}}(t) = \frac{\sum_{j < t-L_t} \frac{\alpha_j}{1-\lambda_j} \mathbf{1}\{P_j > \lambda_j\} + \sum_{j=t-L_t}^{t} \frac{\alpha_j}{1-\lambda_j}}{(\sum_{j \le t, j \notin \{t-L_t,\dots,t-1\}} R_j) \vee 1}.$$

In the case of running asynchronous tests, the algorithms were constructed as pessimistic; however, they had access to as much information as the statistician performing the tests. Here, that is not the case—LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ choose to ignore the outcomes of completed tests as long as they are in the conflict set of subsequent tests. Only after the last-conflict time τ_i, positive outcomes are rewarded by readjusting the FDP estimate. On the other hand, the statistician's perspective is different—as soon as round t is over, the statistician knows the outcome of the t-th test. Just like testing in parallel, testing locally dependent p-values comes at a cost—if the lags are large, the algorithm keeps increasing the FDP estimate, assigning ever smaller test levels, waiting for rewards from tests performed a long time ago. In the extreme case of $L_t = t$, the test levels steadily decrease so that their sum converges to α, regardless of the fact that discoveries have possibly been made.

Explicit setting-specific algorithms, obtained by substituting $\mathcal{X}^t$ for $\mathcal{X}_{\text{dep}}^t$ in Algorithms 1-4, resulting in LORD++, LOND, SAFFRON, and alpha-investing under local dependence, are given in the Appendix.

5. Example 3: Controlling FDR in asynchronous mini-batch testing

Here we merge the ideas of the previous two sections, bringing together asynchronous testing and local dependence of p-values. Although there are various ways one could think of in which these two concepts intertwine, here we discuss a particularly simple and natural one.

Let a *mini-batch* represent a grouping of an arbitrary number of tests that are run asynchronously, which result in dependent p-values; for instance, these tests could be run on the same data. After a mini-batch of tests is fully executed, a new one can start, testing new hypotheses, independent of the previous batch, and doing so on fresh data. From the point of view of asynchrony, such a process could be thought of as a compromise between synchronous and asynchronous testing—batches are internally asynchronous, however they are globally sequential and synchronous. If all batches are of size one, one recovers classical online

testing; if the batch-size tends to infinity, the usual notion of asynchronous testing is obtained. Figure 4 depicts an example of a mini-batch testing process with three mini-batches.

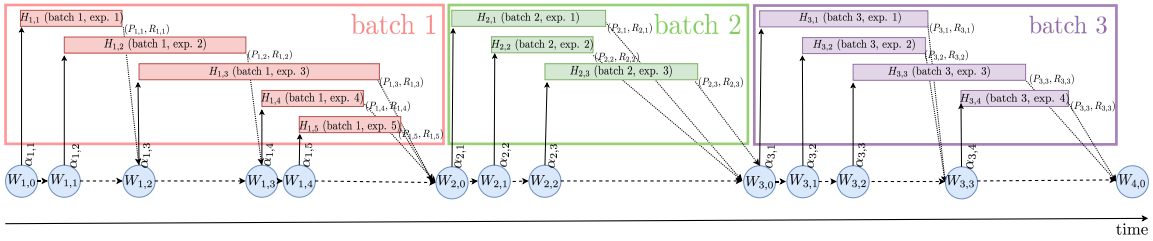

Figure 4: Running three mini-batches of tests. The batches are run synchronously, while the tests that comprise each of them are run asynchronously. We use $W_{t,j-1}$ to denote the remaining "wealth" for making false discoveries before starting the j-th test in the t-th batch.

We introduce notation that captures this setting. We will use two time indices; $P_{b,t}$ denotes the p-value resulting from the test that starts as the t-th one in the b-th batch, testing hypothesis $H_{b,t}$. We allow any two p-values in the same batch to have arbitrary dependence; however, we require any two p-values in different batches to be independent. This can be written compactly as:

$$P_{b_1,i} \perp P_{b_2,j}, \text{ for any } b_1, b_2, i, j, \text{ such that } b_1 \neq b_2.$$

We will denote the size of the b-th batch as n_b. Thus, the first batch results in $P_{1,1}, \ldots, P_{1,n_1}$, the second one in $P_{2,1}, \ldots, P_{2,n_2}$, etc. Analogously, the test levels and candidacy thresholds will also be doubly-indexed; $\alpha_{b,t}$ and $\lambda_{b,t}$ are used for testing $P_{b,t}$. Further, we define $R_{b,t} := \mathbf{1}\{P_{b,t} \leq \alpha_{b,t}\}$, and $C_{b,t} := \mathbf{1}\{P_{b,t} \leq \lambda_{b,t}\}$ as the rejection and candidacy indicators, respectively. By $\mathcal{R}_b$ we will denote the set of rejections in the b-th batch, and by $\mathcal{C}_b$ the set of candidates in the b-th batch.

Recall the key ideas of the previous two sections—tests running in parallel, or those resulting in dependent p-values, are seen as conflicting. We again pursue this approach, and let the conflict set of $P_{b,t}$ consist of all other p-values in the same batch. More formally, the mini-batch conflict set can be defined as:

$$\mathcal{X}^{b,t}_{\text{mini}} = \{(b, i) : i < t\}.$$

Notice that in Section 3, the conflicts arise solely due to missing information, in Section 4 solely due to dependence, while here they are due to both.

The instances of LORD* and SAFFRON* used to test mini-batches will be referred to as LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$. As before, we will define the past-describing filtrations for both of these algorithms. Due to local dependence, as in Section 4, whole batches of tests are mutually conflicted. Only at the finish time of a batch are the discoveries taken into account. For this reason, from the perspective of any batch, all rejections in any prior batch happened at one time step. Consequently, there is no need to consider the actual finish time of any test from previous batches, and thus the respective non-conflicting filtrations for LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ will be of the form:

$$\mathcal{L}^{-\mathcal{X}^{b,t}}_{\text{mini}} = \sigma(\mathcal{R}_1, \ldots, \mathcal{R}_{b-1}),\ \mathcal{S}^{-\mathcal{X}^{b,t}}_{\text{mini}} = \sigma(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{b-1}, \mathcal{C}_{b-1}).$$

As before, we use $\mathcal{F}^{-\mathcal{X}^{b,t}}_{\text{mini}}$ to refer to both of these two filtrations simultaneously. The test levels $\{\alpha_{b,t}\}$ and candidacy thresholds $\{\lambda_{b,t}\}$ are therefore computed as functions of the outcomes of the tests in previous batches, i.e., we can write $\alpha_{b,t} = f_{b,t}(\mathcal{R}_1, \ldots, \mathcal{R}_{b-1})$ for LORD$_{\text{mini}}$, and similarly, $\alpha_{b,t} = g_{b,t}(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{b-1}, \mathcal{C}_{b-1})$ and $\alpha_{b,t} = h_{b,t}(\mathcal{R}_1, \mathcal{C}_1, \ldots, \mathcal{R}_{b-1}, \mathcal{C}_{b-1})$ for SAFFRON$_{\text{mini}}$.

By analogy with the last section, we do not necessarily expect the p-value $P_{b,t}$ to be well-behaved, given that we have seen the outcomes of tests whose p-values have dependence on $P_{b,t}$. By the local dependence assumption, it is straightforward to verify that the following condition holds true:

$$\text{If the null hypothesis } H_{b,t} \text{ is true, then } \Pr\left\{P_{b,t} \leq u \;\middle|\; \mathcal{F}^{-\mathcal{X}^{b,t}}_{\text{mini}}\right\} \leq u, \text{ for all } u \in [0, 1]. \tag{8}$$

The LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ algorithms

By definition of the mini-batch conflict set and the general estimate of LORD*, LORD$_{\text{mini}}$ is obtained as an update rule for $\alpha_{b,t}$ such that the following quantity is controlled for all $b, t \in \mathbb{N}$:

$$\widehat{\text{FDP}}_{\text{LORD}_{\text{mini}}}(b,t) = \frac{\sum_{i<b}\sum_{j\le n_i}\alpha_{i,j} + \sum_{j\le t}\alpha_{b,j}}{(\sum_{i<b}\sum_{j\le n_i} R_{i,j}) \vee 1}.$$

Similarly, SAFFRON$_{\text{mini}}$ controls the following adaptive estimate:

$$\widehat{\text{FDP}}_{\text{SAFFRON}_{\text{mini}}}(b,t) = \frac{\sum_{i<b}\sum_{j\le n_i}\frac{\alpha_{i,j}}{1-\lambda_{i,j}}\mathbf{1}\{P_{i,j} > \lambda_{i,j}\} + \sum_{j\le t}\frac{\alpha_{b,j}}{1-\lambda_{b,j}}}{(\sum_{i<b}\sum_{j\le n_i} R_{i,j}) \vee 1}.$$

Since the set of rejections corresponding to tests that are not in the current conflict set is invariant throughout the testing of any whole batch, the FDP estimate gradually increases while a batch is being tested. Only when the batch has finished testing in its entirety does the algorithm get rewarded for every rejection it made in that batch. This implies that the batch size should be carefully chosen, as the achieved power decreases with batch size. This is numerically verified in Section 7.

The LORD++, LOND, SAFFRON, and alpha-investing procedures for mini-batch testing are explicitly stated in the Appendix, obtained by substituting $\mathcal{X}^{b,t}_{\text{mini}}$ into Algorithms 1-4.

6. Controlling mFDR and FDR at fixed and stopping times

The previous three sections have shown that the abstract framework of conflict sets is a useful representational tool for expressing interactions across different tests, yielding three natural specific testing protocols. In this section, we return to the abstract unified framework in order to prove mFDR guarantees of LORD* and SAFFRON*, which implies mFDR control of all of the setting-specific algorithms. Additionally, we provide several results on strict FDR control under asynchrony and dependence, although under more stringent conditions.

6.1 mFDR control

We begin by focusing on fixed-time mFDR control. As mentioned earlier, the claim for LORD* follows trivially from Proposition 1, so the proof of Theorem 2, given in the Appendix, focuses on providing guarantees for SAFFRON*.

Theorem 2 *Suppose that the null p-values are super-uniform conditional on $\mathcal{F}^{-\mathcal{X}^{E_t}}$, meaning $Pr\Big\{P_t \le u \;\Big|\; \mathcal{F}^{-\mathcal{X}^{E_t}}\Big\} \le u$, for all $u \in [0,1]$ and $t \in \mathcal{H}^0$. Then, LORD* and SAFFRON* with target FDR level α both guarantee that* $\text{mFDR}(t) \le \alpha$ *for all* $t \in \mathbb{N}$.

Notice that the super-uniformity assumption above reduces to conditions (6), (7), and (8), in the three settings previously described.

The result of Theorem 2 actually holds more generally; in particular, in the following theorem we show that mFDR is also controlled at certain stopping times. Our approach is based on constructing a process which behaves similarly to a submartingale, which allows us to derive a result mimicking optional stopping. This process, however, is not a submartingale in the general case. For example, it is not a submartingale in the synchronous setting under local dependence, described in Section 4.

More specifically, we show that LORD* and SAFFRON* control mFDR at any stopping time T which satisfies the following conditions:

(C1) T is defined with respect to the filtration $\mathcal{F}^{-\mathcal{X}^{t+1}}$, $\{T = t\} \in \mathcal{F}^{-\mathcal{X}^{t+1}}$;

(C2) T is almost surely bounded.

Recall that $\mathcal{F}^{-\mathcal{X}^{t+1}}$ denotes the non-conflicting information about the first t tests (in particular, not the first $t+1$), and hence the offset by 1 in indexing. Intuitively, this means that the decision to stop at time t can depend on all information up to time t that the algorithm is allowed to utilize.

Condition (C2) is a mild one, as in practice we primarily care about bounded stopping times. For instance, one would not wait infinitely long to observe the first rejection; if T_{r_1} denotes the time of the first rejection, a natural stopping time would be $T := T_{r_1} \wedge t_{\max}$, where $t_{\max}$ is the fixed longest time one is willing to wait for a rejection.

When $E_t = t$ and the p-values are independent, we require the same conditions as Foster and Stine (2008) in their stopping-time analysis of the mFDR. Consequently, their result can be seen as a special case of Theorem 3, given that alpha-investing is a special instance of SAFFRON.

Theorem 3 *Suppose that the null p-values are super-uniform conditional on $\mathcal{F}^{-\mathcal{X}^{E_t}}$, meaning $Pr\Big\{P_t \le u \;\Big|\; \mathcal{F}^{-\mathcal{X}^{E_t}}\Big\} \le u$, for all $u \in [0,1]$ and $t \in \mathcal{H}^0$. Consider any stopping time T that satisfies conditions (C1-C2). Then, LORD* and SAFFRON* with target FDR level α both control* mFDR *at T:* $\mathrm{mFDR}(T) \le \alpha$.

6.2 FDR control

Even though the main objective of the paper is to provide mFDR guarantees, one can also obtain FDR control for LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$, provided that the p-values in the sequence are independent. This is in line with earlier work where (synchronous) online FDR control has only been proved under independence assumptions (Javanmard and Montanari, 2018; Ramdas et al., 2017, 2018). While our arguments below generalize the earlier ones, we stress that the independence assumption may not be reasonable in asynchronous settings, which is why we focused on the mFDR for most of the paper and we only present the argument below for completeness.

For FDR control, we additionally require α_t and λ_t to be *monotone*. In the context of LORD$_{\text{async}}$, this means that

$$\alpha_t = f_t(\mathcal{R}_1, \ldots, \mathcal{R}_i, \ldots \mathcal{R}_{t-1}) \ge f_t(\mathcal{R}_1, \ldots, \mathcal{R}'_i, \ldots \mathcal{R}_{t-1}) = \alpha'_t$$

whenever $\mathcal{R}'_i \subseteq \mathcal{R}_i$. For SAFFRON$_{\text{async}}$, we require the same condition also when $\mathcal{C}'_i \subseteq \mathcal{C}_i$, both for α_t and λ_t. All update rules stated in this paper are monotone by design.

First we state a technical lemma that is the key ingredient in proving FDR control of our asynchronous procedures, which generalizes several similar lemmas that have appeared in related work (Javanmard and Montanari, 2018; Ramdas et al., 2017, 2018).

Lemma 4 *Assume that null p-values are independent of each other and of the non-nulls. Moreover, let $g : \{\mathbb{N} \cup \{0\}\}^M \to \mathbb{R}$ be any coordinate-wise non-decreasing function. Then, for any index $t \le M$ such that $t \in \mathcal{H}^0$, we have:*

$$\mathbb{E}\left[\frac{\alpha_t \mathbf{1}\{P_t > \lambda_t\}}{(1-\lambda_t) g(|\mathcal{R}|_{1:M})} \;\middle|\; \mathcal{F}^{-\mathcal{X}^{E_t}}_{\textit{async}}\right] \ge \mathbb{E}\left[\frac{\alpha_t}{g(|\mathcal{R}|_{1:M})} \;\middle|\; \mathcal{F}^{-\mathcal{X}^{E_t}}_{\textit{async}}\right] \ge \mathbb{E}\left[\frac{\mathbf{1}\{P_t \le \alpha_t\}}{g(|\mathcal{R}|_{1:M})} \;\middle|\; \mathcal{F}^{-\mathcal{X}^{E_t}}_{\textit{async}}\right],$$

where $|\mathcal{R}|_{1:M} = (|\mathcal{R}_1|, \ldots, |\mathcal{R}_M|)$.

With this lemma, we directly obtain FDR guarantees of LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ under independence, as stated in Theorem 5.

Theorem 5 *Suppose that the null p-values are independent of each other and of the non-nulls, and that α_t and λ_t are monotone. Then, LORD$_{\textit{async}}$ and SAFFRON$_{\textit{async}}$ with target FDR level α both guarantee* $\mathrm{FDR}(t) \le \alpha$ *for all $t \in \mathbb{N}$.*

Additionally, we prove that the original LOND algorithm (Javanmard and Montanari, 2015) controls FDR for an arbitrary sequence of p-values that satisfy positive regression dependency on a subset (PRDS) (Benjamini and Yekutieli, 2001), without any correction. In other words, under the PRDS assumption, it

suffices to take all conflict sets in the sequence to be empty. For convenience, we state the formal definition of PRDS in the Appendix.

Recall the setup of the LOND algorithm. Given a non-negative sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty}\gamma_j = 1$, the test levels are set as $\alpha_t = \alpha\gamma_t(|\mathcal{R}(t-1)| \vee 1)$, where $|\mathcal{R}(t-1)|$ denotes the number of rejections at time $t-1$. Note that this rule is monotone, in the sense that α_t is coordinate-wise non-decreasing in the vector of rejection indicators $(R_1, \dots, R_{t-1})$. Below, we prove that LOND controls the FDR at any time $t \in \mathbb{N}$ under PRDS.

Recalling the definition of reshaping (Ramdas et al., 2019; Blanchard and Roquain, 2008), we will also prove that if $\{\beta_t\}$ is a sequence of reshaping functions, then using the test levels $\widetilde{\alpha}_t := \alpha\gamma_t\beta_t(|\mathcal{R}(t-1)| \vee 1)$ controls FDR under arbitrary dependence. We call this the reshaped LOND algorithm. As one example, using the Benjamini-Yekutieli reshaping yields $\widetilde{\alpha}_t := \alpha\gamma_t(|\mathcal{R}(t-1)| \vee 1)/(\sum_{i=1}^{t}\frac{1}{i})$.

Theorem 6 *(a) The LOND algorithm satisfies* $\text{FDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$ *under positive dependence (PRDS).*

(b) Reshaped LOND satisfies $\text{FDR}(t) \leq \alpha$ *for all* $t \in \mathbb{N}$ *under arbitrary dependence.*

7. Numerical experiments

Here we present the results of several numerical simulations, which show the gradual change in performance of LORD* and SAFFRON* with the increase of asynchrony and the lags of local dependence.[3] We also compare these solutions to existing procedures with formal FDR guarantees under dependence. The plots in this section compare the achieved power and FDR of $\text{LORD}_{\text{async}}$, $\text{SAFFRON}_{\text{async}}$, LORD_{dep}, $\text{SAFFRON}_{\text{dep}}$, $\text{LORD}_{\text{mini}}$ and $\text{SAFFRON}_{\text{mini}}$ for different problem parameters, in settings with p-values computed from Gaussian observations. We present additional experiments, including those on real data, in the Appendix.

The justification for focusing on synthetic data is two-fold. First, there is no standardized real data set for testing online FDR procedures. The quintessential applications of these methods involve testing with sensitive data, which are not publicly available due to privacy concerns. Second, even when real data are obtainable, it is unclear how one would evaluate the ground truth.

In all of the simulations we present the FDR is controlled at $\alpha = 0.05$, and we estimate the FDR and power by averaging the results of 200 independent trials. The SAFFRON-type algorithms use the constant candidacy threshold sequence $\lambda = 1/2$, across all tests. The LORD-type algorithms use the LORD++ update for test levels. Each figure additionally plots the performance of uncorrected testing, in which the constant test level $\alpha_t = \alpha = 0.05$ is used across all $t \in \mathbb{N}$, and alpha-spending, whose test levels decay according to the $\{\gamma_t\}_{t=1}^{\infty}$ sequence of LORD* and SAFFRON*.

The experiments test for the means of $M = 1000$ Gaussian observations, and each null hypothesis takes the form $H_i : \mu_i = 0$, where μ_i is the mean of the Gaussian sample. We generate samples $\{Z_i\}_{i=1}^{M}$, where $Z_i \sim N(\mu_i, 1)$ and the parameter μ_i is chosen as $\mu_i = \xi F_1$, where $\xi \sim \text{Bern}(\pi_1)$, for a fixed proportion of non-nulls in the sequence π_1, and some random variable F_1. We consider two distributions for F_1—a degenerate distribution with a point mass at μ_c, where μ_c is a fixed constant for the whole sequence, or $N(0, 2\log(M))$. The motivation for the latter is that $\sqrt{2\log(M)}$ is the minimax amplitude for estimation under the sparse Gaussian sequence model. In the case of the mean coming from a degenerate distribution, we form one-sided p-values as $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian CDF. If the mean has a Gaussian distribution, we form two-sided p-values, i.e., $P_i = 2\Phi(-|Z_i|)$.

7.1 Varying asynchrony

First we show the results of simulated asynchronous tests, in which the p-values are independent. At each time step, the test duration is sampled from a geometric distribution with parameter p: $E_i \sim i - 1 + \text{Geom}(p)$ for all i. This implies that $p = 1$ yields the fully synchronous setting, while, as p gets smaller, the expectation of

3. The code for all experiments in this section is available at: https://github.com/tijana-zrnic/async-online-FDR-code

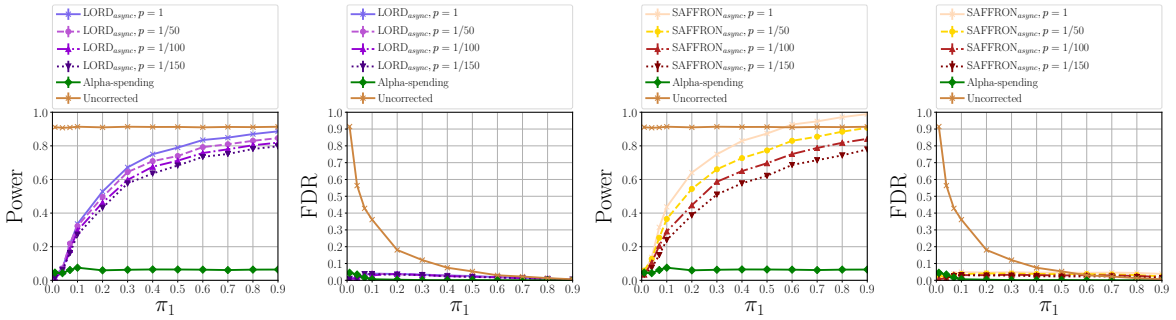

Figure 5: Power and FDR of LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ with varying the parameter of asynchrony p of the tests. In all five runs LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is a point mass at $\mu_c = 3$.

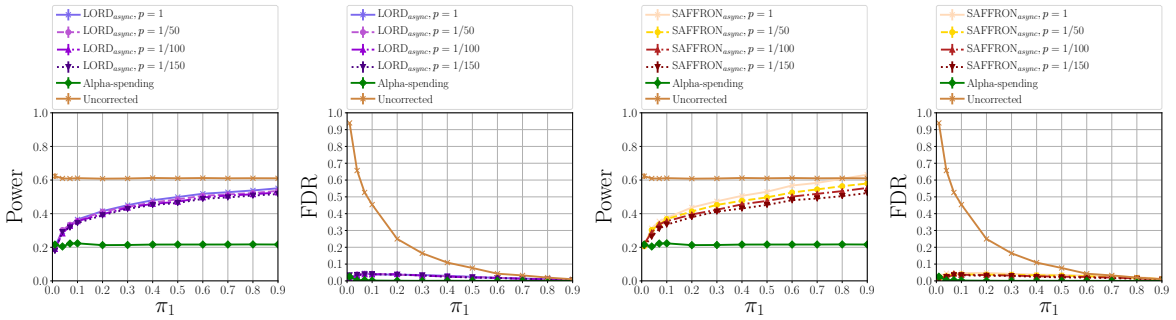

Figure 6: Power and FDR of LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ with varying the parameter of asynchrony p of the tests. In all five runs LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is $N(0, 2\log(M))$.

the test duration grows larger, hence the procedure gets more asynchronous, and consequently less powerful. Figure 5 shows numerically how changing p affects the achieved power of LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$, across different non-null proportions π_1, when the mean of the alternative is fixed as $\mu_c = 3$. Figure 6 plots power and FDR of LORD$_{\text{async}}$ and SAFFRON$_{\text{async}}$ against π_1 for normally distributed means, showing a more gradual change in performance with the increase of asynchrony.

Building on our discussion in Section 1.2, we note that certain sequential tests have simpler, less "optimized" counterparts that do not require knowledge of the test level up front. As a result, standard online FDR algorithms can be applied. This leads to another tradeoff in asynchronous testing—one between the power gain of optimized tests, which make use of the test level at the beginning of the test, and the accompanying power loss when running such tests asynchronously. We illustrate this point numerically in Appendix D, where we simulate A/B tests of varying levels of asynchrony using two approaches: an optimized one which makes use of α_i throughout the test, and a naive one which computes a p-value without knowledge of α_i.

7.2 Varying the lag of dependence

The second set of simulations considers synchronous testing of locally dependent p-values. We take L_t to be invariant and equal to L, which reduces to lagged dependence. We generate an M-dimensional vector of Gaussian observations $(Z_1, \ldots, Z_M)$, which are marginally distributed according to the model described at

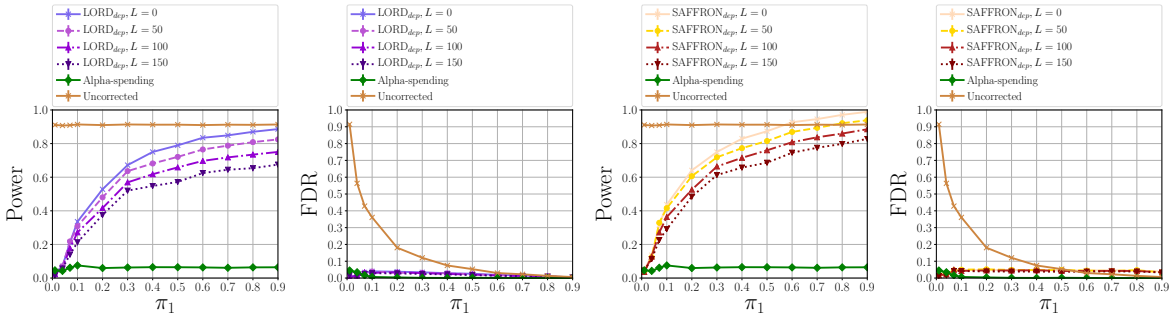

Figure 7: Power and FDR of LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ with varying the dependence lag L in the p-value sequence. In all five runs LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is a point mass at $\mu_c = 3$.

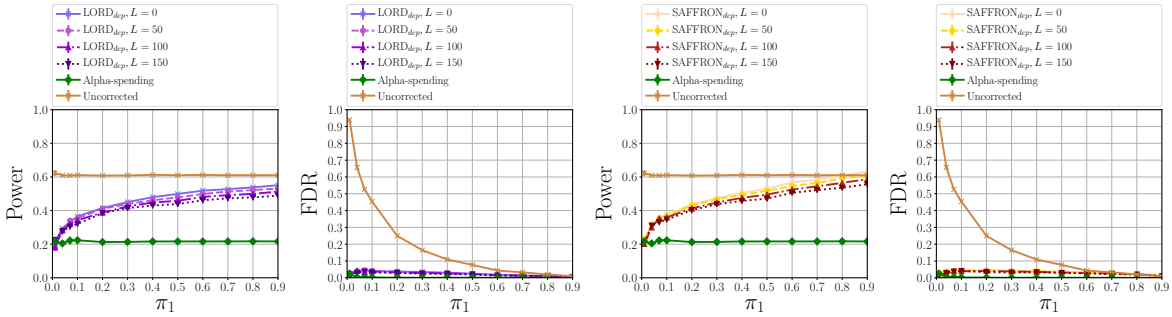

Figure 8: Power and FDR of LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ with varying the dependence lag L in the p-value sequence. In all five runs LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is $N(0, 2\log(M))$.

the beginning of the section, and have the following $M \times M$ Toeplitz covariance matrix:

$$\Sigma(M, L, \rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^L & 0 & \dots & 0 & 0 & 0 \\ \rho & 1 & \rho & \dots & \rho^{L-1} & \rho^L & \dots & 0 & 0 & 0 \\ \vdots & & & \ddots & & & & & \vdots & \\ \vdots & & & & \ddots & & & & \vdots & \\ \vdots & & & & & \ddots & & & \vdots & \\ \vdots & & & & & & \ddots & & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \rho & 1 & \rho \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \rho^2 & \rho & 1 \end{bmatrix}, \tag{9}$$

where we set $\rho = 0.5$. Figure 7 compares the power and FDR of LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ under local dependence, when the mean of the observations under the alternative is $\mu_c = 3$ with probability 1. Figure 8 gives the same comparison when the mean of non-null samples is normally distributed, which yields a slower decrease in performance with increasing the lag.

7.3 Varying mini-batch sizes

Here we analyze the change in performance of LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ when the size of mini-batches varies. We fix the batch size $n_b \equiv n$ for all batches b. Within each batch tests are performed asynchronously,

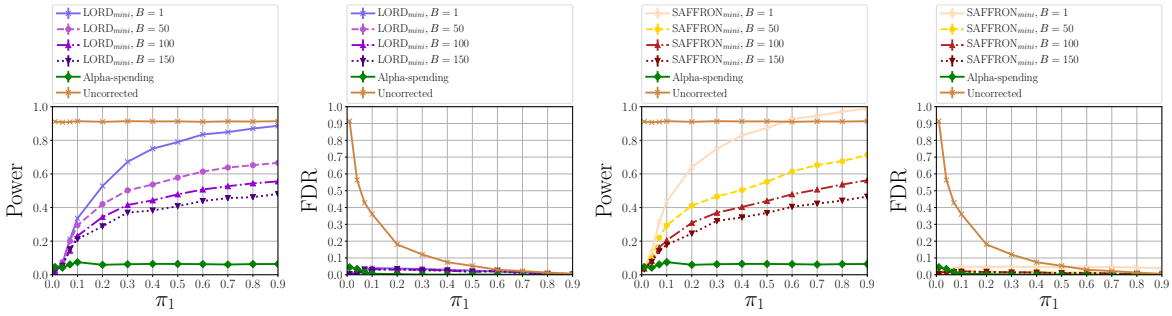

Figure 9: Power and FDR of LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ with varying the size of mini-batches. In all five runs LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is a point mass at $\mu_c = 3$, and $\rho = 0.5$.

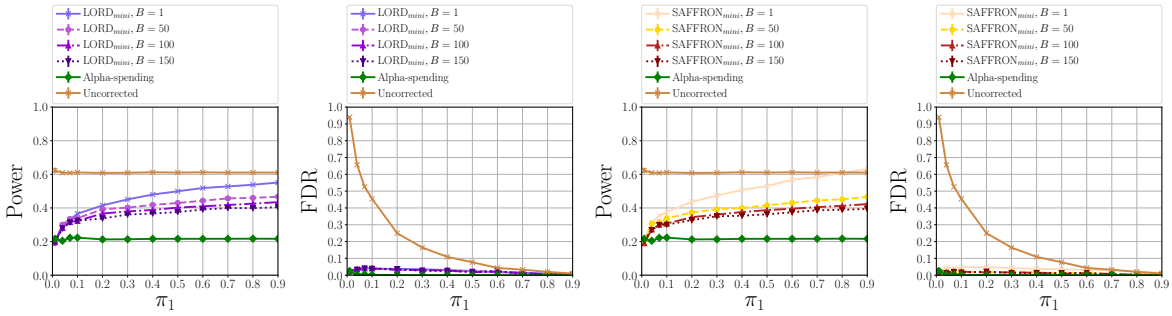

Figure 10: Power and FDR of LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ with varying the size of mini-batches. In all five runs LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ have the same parameters $(\{\gamma_j\}_{j=1}^{\infty}, W_0)$. The mean of observations under the alternative is $N(0, 2\log(M))$, and $\rho = 0.5$.

and all p-values within the same batch are dependent. In particular, they follow a multivariate normal distribution, where the marginal distributions are as described at the beginning of this section, and the covariance matrix is the Toeplitz matrix $\Sigma(n, n-1, \rho)$ (9), where we fix $\rho = 0.5$. Dependent p-values come in "blocks" of size n, implying that any two p-values belonging to two different batches are independent. Figure 9 compares the power and FDR of LORD$_{\text{mini}}$ and SAFFRON$_{\text{mini}}$ for different batch sizes when the mean of the non-null Z_i is a point mass at $\mu_c = 3$, and Figure 10 plots the same comparison when the mean of the non-null observations is normally distributed.

7.4 Comparison with LORD under dependence

The final set of experiments contrasts LORD$_{\text{dep}}$ and SAFFRON$_{\text{dep}}$ to the original LORD algorithm under dependence. The latter controls FDR under arbitrary dependence, however, as mentioned earlier, this entails a similar update to alpha-investing; more precisely, the test levels α_j^{indep} of LORD under independence have to be discounted by a convergent sequence $\{\xi_j\}_{j=1}^{\infty}$, resulting in new test levels $\alpha_j := \xi_j \alpha_j^{\text{indep}}$, which essentially diminishes the effect of α_j^{indep} earning extra budget through discoveries. We generate the p-value sequence using the same scheme as in Subsection 7.2; they are computed from Gaussian observations with covariance matrix $\Sigma(M, L, \rho)$ (9), where we fix $\rho = 0.5$ and $L = 150$. By construction, this sequence is only locally dependent, which implies that the application of our algorithms comes with provable guarantees. Figure 11 compares the power and FDR of SAFFRON$_{\text{dep}}$, LORD$_{\text{dep}}$, LORD under dependence, and alpha-

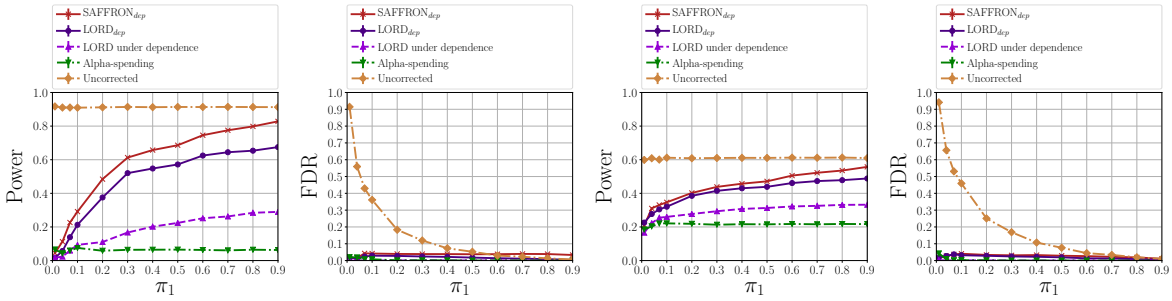

Figure 11: Power and FDR of SAFFRON$_{\text{dep}}$, LORD$_{\text{dep}}$, LORD under dependence, and alpha-spending. The decay of test levels in alpha-spending and discount sequence $\{\xi_j\}_{j=1}^{\infty}$ act according to the sequence $\{\gamma_j\}_{j=1}^{\infty}$ used for SAFFRON$_{\text{dep}}$ and LORD$_{\text{dep}}$. On the left two plots, the mean of observations under the alternative is a point mass at $\mu_c = 3$, while on the right two plots, it is distributed as $N(0, 2\log(M))$. We fix parameters $\rho = 0.5$ and $L = 150$.

spending when the mean of the non-null Z_i is a point mass at $\mu_c = 3$ (left), as well as in the setting with a normally distributed mean under the alternative (right).

8. Discussion

We have presented a unified framework for the design and analysis of online FDR procedures for asynchronous testing, as well as testing locally dependent p-values. Our framework reposes on the concept of "conflict sets," and we show the value of this concept for the study of both asynchronous testing and local dependence and for their combination. We derive two specific procedures that make use of conflict sets to yield algorithms that provide online mFDR and FDR control.

Several technical questions remain open for future work. While we have shown strict FDR control of our asynchronous procedures under independence, it is still unclear how to prove their FDR control under local dependence. We believe that it might also be possible to prove FDR control of uncorrected LORD under positive dependence, similarly to how we proved validity of the plain LOND algorithm under positive dependence in Section 6. Finally, it would be of great interest to obtain strict FDR control at stopping times, a problem that remains open even under independence of p-values. In mFDR control, this proof relies on a martingale-like argument which decouples the numerator and denominator of the FDP. The expected numerator increments, conditional on the information from past tests, are then controlled by invoking super-uniformity. When false rejection indicators are coupled with the FDP denominator, however, it is less clear how to invoke conditional super-uniformity. This is a non-trivial step even when analyzing FDR at fixed times, as witnessed by many "super-uniformity lemmas" in the literature (Javanmard and Montanari, 2018; Ramdas et al., 2017, 2018).

Appendix A. Deferred proofs

A.1 Proof of Proposition 1

Fix a time step $t \in \mathbb{N}$. By this time, exactly t tests have started, and hence at most those t decisions are known. Therefore, by linearity of expectation:

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

Applying the law of iterated expectations by conditioning on $\mathcal{F}^{-\mathcal{X}^{E_j}}$ for each term, we obtain:

$$\sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\mathbf{1}\left\{P_j \le \alpha_j\right\}\right] = \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{P_j \le \alpha_j\right\} \,\middle|\, \mathcal{F}^{-\mathcal{X}^{E_j}}\right]\right] \le \sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\alpha_j\right],$$

which follows due to measurability of α_j with respect to $\mathcal{F}^{-\mathcal{X}^j} \subseteq \mathcal{F}^{-\mathcal{X}^{E_j}}$, and the super-uniformity assumption. If we assume $\text{FDP}^*(t) := \frac{\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j}{(\sum_{E_j \le t} R_j) \vee 1} \le \alpha$, then it follows that:

$$\sum_{j \le t, j \in \mathcal{H}^0} \mathbb{E}\left[\alpha_j\right] = \mathbb{E}\left[\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j\right] \le \alpha \mathbb{E}\left[\left(\sum_{E_j \le t} R_j\right) \vee 1\right] = \alpha \mathbb{E}\left[|\mathcal{R}(t)| \vee 1\right],$$

which follows by linearity of expectation and the assumption on $\text{FDP}^*(t)$. Rearranging yields the inequality $\text{mFDR}(t) := \frac{\mathbb{E}[|\mathcal{V}(t)|]}{\mathbb{E}[|\mathcal{R}(t)| \vee 1]} \le \alpha$, which completes the proof.

A.2 Proof of Theorem 2

As stated before, the guarantees for LORD* follow directly from Proposition 1, after observing that $\text{FDP}^*_{\text{conf}}(t) \le \widehat{\text{FDP}}_{\text{LORD*}}(t) \le \alpha$ holds almost surely for all $t \in \mathbb{N}$. Therefore, in the rest of this proof, we focus on SAFFRON*.

Fix a time t. Then, we have:

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

where the inequality follows because the set of rejections made by time t could be at most the set $[t]$. Note that α_j and λ_j are measurable with respect to $\mathcal{S}^{-\mathcal{X}^j} \subseteq \mathcal{S}^{-\mathcal{X}^{E_j}}$; therefore, applying iterated expectations by conditioning on $\mathcal{S}^{-\mathcal{X}^{E_j}}$ gives:

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

where we apply the super-uniformity assumption. If we assume that

$$\widehat{\text{FDP}}_{\text{SAFFRON*}}(t) := \frac{\sum_{j < t, j \notin \mathcal{X}^t} \frac{\alpha_j}{1-\lambda_j} \mathbf{1}\left\{P_j > \lambda_j\right\} + \sum_{j \in \mathcal{X}^t \cup \{t\}} \frac{\alpha_j}{1-\lambda_j}}{\left(\sum_{j < t, j \notin \mathcal{X}^t} R_j\right) \vee 1} \le \alpha,$$

then it follows that:

$$\begin{aligned}
&\sum_{j\le t, j\in\mathcal{H}^0} \mathbb{E}\left[\alpha_j \frac{\mathbf{1}\{P_j>\lambda_j\}}{1-\lambda_j}\right] \le \sum_{j\le t} \mathbb{E}\left[\alpha_j \frac{\mathbf{1}\{P_j>\lambda_j\}}{1-\lambda_j}\right] \\
&\le \mathbb{E}\left[\sum_{j<t, j\notin\mathcal{X}^t} \frac{\alpha_j}{1-\lambda_j}\mathbf{1}\{P_j>\lambda_j\} + \sum_{j\in\mathcal{X}^t\cup\{t\}} \frac{\alpha_j}{1-\lambda_j}\right] \le \alpha\mathbb{E}\left[\left(\sum_{j<t, j\notin\mathcal{X}^t} R_j\right)\vee 1\right] \\
&\le\ \alpha\mathbb{E}\left[|\mathcal{R}(t)|\vee 1\right],
\end{aligned}$$

where the first inequality drops the condition $j\in\mathcal{H}^0$, the second one ignores the condition $\mathbf{1}\{P_j>\lambda_j\}$ for some terms, the third inequality applies the assumption on $\widehat{\text{FDP}}_{\text{SAFFRON}^*}(t)$ and the last inequality uses the fact that $\mathcal{R}(t)$ contains all past rejections that are no longer conflicting. Rearranging the terms in the previous derivation, we reach the conclusion that $\text{mFDR}(t)\le\alpha$, which concludes the proof of the theorem.

A.3 Proof of Theorem 3

We first prove the theorem for LORD*, and then we move on to proving the SAFFRON* guarantees.

LORD*. For all $t\in\mathbb{N}$, define the process $A(t)$ as:

$$A(t) := -\sum_{i\le t, i\in\mathcal{H}^0} \mathbf{1}\{E_i\le t\}\left(\mathbf{1}\{P_i\le\alpha_i\}-\alpha_i\right) = A(t-1) - \sum_{i\le t, i\in\mathcal{H}^0} \mathbf{1}\{E_i=t\}\left(\mathbf{1}\{P_i\le\alpha_i\}-\alpha_i\right),$$

where we take $A(0)=0$. Let $H(t) := \mathbf{1}\{T\ge t\}$. Since T is a stopping time, it holds that $\{T\ge t+1\} = \{T\le\ t\}^c\in\mathcal{F}^{-\mathcal{X}^{t+1}}$, therefore $H(t+1)$ is predictable, that is it is measurable with respect to $\mathcal{F}^{-\mathcal{X}^{t+1}}$. Define the transform $(H\cdot A)$ of H by A as follows:

$$\begin{aligned}
(H\cdot A)(t) &:= \sum_{m=1}^{t} H(m)(A(m)-A(m-1)) \\
&= \sum_{m=1}^{t} H(m)\left(-\sum_{i\le m, i\in\mathcal{H}^0} \mathbf{1}\{E_i=m\}\left(\mathbf{1}\{P_i\le\alpha_i\}-\alpha_i\right)\right).
\end{aligned}$$

By taking conditional expectations, we can obtain:

$$\begin{aligned}
&\mathbb{E}\left[(H\cdot A)(t+1)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right] \\
&= \mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right] + \mathbb{E}\left[H(t+1)(A(t+1)-A(t))\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right] \\
&= \mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right] \\
&\quad + H(t+1)\mathbb{E}\left[-\sum_{i\le t+1, i\in\mathcal{H}^0} \mathbf{1}\{E_i=t+1\}\left(\mathbf{1}\{P_i\le\alpha_i\}-\alpha_i\right)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right] \\
&= \mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{E_t}}\right] \\
&\quad + H(t+1)\sum_{i\le t+1, i\in\mathcal{H}^0} \mathbf{1}\{E_i=t+1\}\,\mathbb{E}\left[-(\mathbf{1}\{P_i\le\alpha_i\}-\alpha_i)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right],
\end{aligned}$$

where the first and last equality follow by linearity of expectation, and the second one uses the predictability of $H(t+1)$. Now we can apply the super-uniformity condition (3), since we are summing over null indices: $\mathbb{E}\left[-\mathbf{1}\{P_i\le\alpha_i\}+\alpha_i\,\middle|\,\mathcal{F}^{-\mathcal{X}^{E_i}}\right]\ge -\alpha_i+\alpha_i = 0$. Therefore, additionally applying the law of iterated

expectations, it follows that $\mathbb{E}\left[(H \cdot A)(t+1)\right] \geq \mathbb{E}\left[(H \cdot A)(t)\right]$. Iteratively applying the same argument, we reach the conclusion that, for all $t \in \mathbb{N}$:

$$\mathbb{E}\left[(H \cdot A)(t)\right] \geq 0. \tag{10}$$

So far we have only used the predictability of $H(t)$; observe that, by its definition, and the definition of $A(t)$, $(H \cdot A)(t) = A(T \wedge t) - A(0) = A(T \wedge t)$, and hence by equation (10), we obtain $\mathbb{E}\left[(H \cdot A)(t)\right] = \mathbb{E}\left[A(T \wedge t)\right] \geq 0$.

Since $A(T \wedge t) \to A(T)$ almost surely as $t \to \infty$, by boundedness of T and dominated convergence we can conclude that $\mathbb{E}\left[A(T \wedge t)\right] \to \mathbb{E}\left[A(T)\right]$ as $t \to \infty$. With this we obtain a useful intermediate result:

$$\mathbb{E}\left[A(T)\right] \geq 0. \tag{11}$$

Recall that $\mathcal{R}(t)$ denotes the set of all rejections made by time t, and $\mathcal{V}(t)$ denotes the set of false rejections made by time t. Consider the following process:

$$\begin{aligned} B(t) &:= \alpha(|\mathcal{R}(t)| \vee 1) - |\mathcal{V}(t)| + \sum_{j \leq t} \alpha_j - \alpha \left(\sum_{j < t, j \notin \mathcal{X}^t} R_j \vee 1 \right) \\ &\geq -|\mathcal{V}(t)| + \sum_{j \leq t} \alpha_j \geq A(t), \end{aligned}$$

where the final inequality applies the definition of $A(t)$ together with the fact that $\sum_{j \leq t} \alpha_j \geq \sum_{j \leq t} \mathbf{1}\left\{E_j \leq t\right\} \alpha_j$.

Now take a stopping time T such that the conditions of the theorem are satisfied, then:

$$\begin{aligned} \mathbb{E}\left[\alpha(|\mathcal{R}(T)| \vee 1) - |\mathcal{V}(T)|\right] &= \mathbb{E}\left[B(T) - \sum_{j \leq T} \alpha_j + \alpha \left(\sum_{j < T, j \notin \mathcal{X}^T} R_j \vee 1 \right) \right] \\ &\geq \mathbb{E}\left[B(T)\right] \geq \mathbb{E}\left[A(T)\right] \geq 0, \end{aligned}$$

where the first inequality follows by definition of the LORD* FDP estimate, the second one by the relationship already established between $A(t)$ and $B(t)$, and the third inequality applies the intermediate result (11). Rearranging the terms we have that $\text{mFDR}(T) \leq \alpha$, as desired.

SAFFRON*. We begin the proof using similar tools as in the LORD* section of the proof. For all $t \in \mathbb{N}$, define the process $A(t)$ as:

$$\begin{aligned} A(t) &:= - \sum_{i \leq t, i \in \mathcal{H}^0} \mathbf{1}\left\{E_i \leq t\right\} \left(\mathbf{1}\left\{P_i \leq \alpha_i\right\} - \mathbf{1}\left\{P_i > \lambda_i\right\} \frac{\alpha_i}{1 - \lambda_i} \right) \\ &= A(t-1) - \sum_{i \leq t, i \in \mathcal{H}^0} \mathbf{1}\left\{E_i = t\right\} \left(\mathbf{1}\left\{P_i \leq \alpha_i\right\} + \mathbf{1}\left\{P_i > \lambda_i\right\} \frac{\alpha_i}{1 - \lambda_i} \right), \end{aligned}$$

where we take $A(0) = 0$. Let $H(t) := \mathbf{1}\left\{T \geq t\right\}$. Since T is a stopping time, it holds that $\{T \geq t+1\} = \{T \leq t\}^c \in \mathcal{F}^{-\mathcal{X}^{t+1}}$, therefore $H(t+1)$ is measurable with respect to $\mathcal{F}^{-\mathcal{X}^{t+1}}$. Define the following transform of H by A:

$$\begin{aligned} (H \cdot A)(t) &:= \sum_{m=1}^{t} H(m)(A(m) - A(m-1)) \\ &= \sum_{m=1}^{t} H(m) \left(- \sum_{i \leq m, i \in \mathcal{H}^0} \mathbf{1}\left\{E_i = m\right\} \left(\mathbf{1}\left\{P_i \leq \alpha_i\right\} + \mathbf{1}\left\{P_i > \lambda_i\right\} \frac{\alpha_i}{1 - \lambda_i} \right) \right). \end{aligned}$$

By taking conditional expectations, we can obtain:

$$\begin{aligned}
&\mathbb{E}\left[(H\cdot A)(t+1)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]\\
&=\mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]+\mathbb{E}\left[H(t+1)(A(t+1)-A(t))\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]\\
&=\mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]\\
&\quad+H(t+1)\mathbb{E}\left[-\sum_{i\le t+1,i\in\mathcal{H}^0}\mathbf{1}\left\{E_i=t+1\right\}\left(\mathbf{1}\left\{P_i\le\alpha_i\right\}+\mathbf{1}\left\{P_i>\lambda_i\right\}\frac{\alpha_i}{1-\lambda_i}\right)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]\\
&=\mathbb{E}\left[(H\cdot A)(t)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right]\\
&\quad+H(t+1)\sum_{i\le t+1,i\in\mathcal{H}^0}\mathbf{1}\left\{E_i=t+1\right\}\mathbb{E}\left[-\left(\mathbf{1}\left\{P_i\le\alpha_i\right\}+\mathbf{1}\left\{P_i>\lambda_i\right\}\frac{\alpha_i}{1-\lambda_i}\right)\,\middle|\,\mathcal{F}^{-\mathcal{X}^{t+1}}\right],
\end{aligned}$$

where the first equality follows by linearity of expectation and the definition of the transform and the second one uses measurability of $H(t+1)$. The term $-\mathbf{1}\left\{E_i=t+1\right\}\left(\mathbf{1}\left\{P_i\le\alpha_i\right\}+\mathbf{1}\left\{P_i>\lambda_i\right\}\frac{\alpha_i}{1-\lambda_i}\right)$ is clearly non-negative when $E_i\neq t+1$. If $E_i=t+1$ however, we can invoke the super-uniformity condition (3), since we are summing over null indices:

$$\mathbb{E}\left[-(\mathbf{1}\left\{P_i\le\alpha_i\right\}+\mathbf{1}\left\{P_i>\lambda_i\right\}\frac{\alpha_i}{1-\lambda_i})\,\middle|\,\mathcal{F}^{-\mathcal{X}^{E_i}}\right]\ge-\alpha_i+(1-\lambda_i)\frac{\alpha_i}{1-\lambda_i}=0.$$

Therefore, additionally applying the law of iterated expectations, it follows that $\mathbb{E}\left[(H\cdot A)(t+1)\right]\ge\mathbb{E}\left[(H\cdot A)(t)\right]$. Iteratively applying the same argument, we reach the conclusion that, for all $t\in\mathbb{N}$:

$$\mathbb{E}\left[(H\cdot A)(t)\right]\ge 0. \tag{12}$$

So far we have only used the predictability of $H(t)$; observe that, by its definition, and the definition of $A(t)$, $(H\cdot A)(t)=A(T\wedge t)-A(0)=A(T\wedge t)$, and hence by equation (12), we obtain $\mathbb{E}\left[(H\cdot A)(t)\right]=\mathbb{E}\left[A(T\wedge t)\right]\ge 0$.

Since $A(T\wedge t)\to A(T)$ almost surely as $t\to\infty$, by boundedness of T and dominated convergence we can conclude that $\mathbb{E}\left[A(T\wedge t)\right]\to\mathbb{E}\left[A(T)\right]$ as $t\to\infty$. As in the LORD* argument, we reach the result that states:

$$\mathbb{E}\left[A(T)\right]\ge 0. \tag{13}$$

Recall $\mathcal{R}(t)$, the set of all rejections made by time t, and $\mathcal{V}(t)$, the set of false rejections made by time t. Consider the following process:

$$\begin{aligned}
B(t)&:=\alpha(|\mathcal{R}(t)|\vee 1)-|\mathcal{V}(t)|+\sum_{j<t,j\notin\mathcal{X}^t}\mathbf{1}\left\{P_j>\lambda_j\right\}\frac{\alpha_j}{1-\lambda_j}+\sum_{j\in\mathcal{X}^t\cup\{t\}}\frac{\alpha_j}{1-\lambda_j}-\left(\sum_{j<t,j\notin\mathcal{X}^t}R_j\vee 1\right)\alpha\\
&\ge-|\mathcal{V}(t)|+\sum_{j\le t}\mathbf{1}\left\{P_j>\lambda_j\right\}\frac{\alpha_j}{1-\lambda_j}\ge A(t),
\end{aligned}$$

where the second inequality applies the definition of $A(t)$ together with $\mathbf{1}\left\{E_j\le t\right\}\le 1$.

Now taking a stopping time T that satisfies the conditions of the theorem, we have:

$$\begin{aligned}
&\mathbb{E}\left[\alpha(|\mathcal{R}(t)|\vee 1)-|\mathcal{V}(t)|\right]\\
&=\mathbb{E}\left[B(T)-\sum_{j<T,j\notin\mathcal{X}^T}\mathbf{1}\left\{P_j>\lambda_j\right\}\frac{\alpha_j}{1-\lambda_j}-\sum_{j\in\mathcal{X}^T\cup\{T\}}\frac{\alpha_j}{1-\lambda_j}+\left(\sum_{j<T,j\notin\mathcal{X}^T}R_j\vee 1\right)\alpha\right]\\
&\ge\mathbb{E}\left[B(T)\right]\ge\mathbb{E}\left[A(T)\right]\ge 0,
\end{aligned}$$

where the first inequality follows by construction of the SAFFRON* empirical FDP estimate, the second inequality uses the proved relationship between $A(t)$ and $B(t)$, and the third inequality applies equation (13). Rearranging the terms we have that mFDR$(T) \leq \alpha$, as desired.

A.4 Proof of Lemma 4

We begin by focusing on the first inequality. Letting $P_{1:M} = (P_1, \ldots, P_M)$ be the original vector of p-values, we define a "hallucinated" vector of p-values $\widetilde{P}^{t\to 1}_{1:M} := (\widetilde{P}_1, \ldots, \widetilde{P}_M)$ that equals $P_{1:M}$, except that the t-th component is set to one:

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

Further, denote by $\widetilde{E}_j$ the finish times of the tests that yield $\widetilde{P}_j$, and let $\widetilde{E}_j$ be equal to E_j for all $1 \leq j \leq M$. Denote the set of candidates and rejections in the hallucinated sequence at time i by $\widetilde{\mathcal{C}}_i$ and $\widetilde{\mathcal{R}}_i$, respectively, and let $\widetilde{\alpha}_i$ be the test level for $\widetilde{P}_i$. Also, let $\mathcal{R}_{1:M} = (\mathcal{R}_1, \ldots, \mathcal{R}_M)$ and $\widetilde{\mathcal{R}}^{t\to 1}_{1:M} = (\widetilde{\mathcal{R}}_1, \ldots, \widetilde{\mathcal{R}}_M)$ denote the vectors of the numbers of rejections using $P_{1:M}$ and $\widetilde{P}^{t\to 1}_{1:M}$, respectively. Similarly, let $\mathcal{C}_{1:M} = (\mathcal{C}_1, \ldots, \mathcal{C}_M)$ and $\widetilde{\mathcal{C}}^{t\to 1}_{1:M} = (\widetilde{\mathcal{C}}_1, \ldots, \widetilde{\mathcal{C}}_M)$ denote the vectors of the numbers of candidates using $P_{1:M}$ and $\widetilde{P}^{t\to 1}_{1:M}$, respectively.

By construction, we have the following properties:

1. $\widetilde{E}_j = E_j, \forall j$ implies $\alpha_i = \widetilde{\alpha}_i$ for all $i \leq E_t$.
2. $\widetilde{\mathcal{R}}_i = \mathcal{R}_i$ and $\widetilde{\mathcal{C}}_i = \mathcal{C}_i$ for all $i < E_t$, since the p-values from the finished tests and the respective test levels are the same in the original and hallucinated setting.
3. $\widetilde{\mathcal{R}}_{E_t} \subseteq \mathcal{R}_{E_t}$ and $\widetilde{\mathcal{C}}_{E_t} \subseteq \mathcal{C}_{E_t}$, and hence $\widetilde{\mathcal{R}}_i \subseteq \mathcal{R}_i$ also for all $i > E_t$, due to monotonicity of the test levels α_i and candidacy thresholds λ_i.

Therefore, on the event $\{P_t > \lambda_t\}$, we have $\mathcal{R}_{E_t} = \widetilde{\mathcal{R}}_{E_t}$ and $\mathcal{C}_{E_t} = \widetilde{\mathcal{C}}_{E_t}$, and hence also $\mathcal{R}_{1:M} = \widetilde{\mathcal{R}}^{t\to 1}_{1:M}$ and $\mathcal{C}_{1:M} = \widetilde{\mathcal{C}}^{t\to 1}_{1:M}$. This allows us to conclude that:

$$\frac{\alpha_t \mathbf{1}\{P_t > \lambda_t\}}{(1-\lambda_t) g(|\mathcal{R}|_{1:M})} = \frac{\alpha_t \mathbf{1}\{P_t > \lambda_t\}}{(1-\lambda_t) g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})}.$$

Since the null p-values are mutually independent and independent of the non-nulls, we conclude that $\widetilde{\mathcal{R}}^{t\to 1}_{1:M}$ is independent of P_t conditioned on $\mathcal{F}^{-\mathcal{X}^{E_t}}_{\text{async}}$. With this, we can obtain:

$$\begin{aligned}\mathbb{E}\left[\frac{\alpha_t \mathbf{1}\{P_t > \lambda_t\}}{(1-\lambda_t) g(|\mathcal{R}|_{1:M})} \,\middle|\, \mathcal{F}^{-\mathcal{X}^{E_t}}_{\text{async}}\right] &= \mathbb{E}\left[\frac{\alpha_t \mathbf{1}\{P_t > \lambda_t\}}{(1-\lambda_t) g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})} \,\middle|\, \mathcal{F}^{-X^{E_t}}_{\text{async}}\right] \geq \mathbb{E}\left[\frac{\alpha_t}{g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})} \,\middle|\, \mathcal{F}^{-\mathcal{X}^{E_t}}_{\text{async}}\right] \\ &\geq \mathbb{E}\left[\frac{\alpha_t}{g(|\mathcal{R}|_{1:M})} \,\middle|\, \mathcal{F}^{-\mathcal{X}^{E_t}}_{\text{async}}\right],\end{aligned}$$

where the first inequality follows by taking an expectation only with respect to P_t by invoking the asynchronous super-uniformity property (6), and the second inequality follows because $g(|\mathcal{R}|_{1:M}) \geq g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})$ since $|\mathcal{R}_i| \geq |\widetilde{\mathcal{R}}_i|$ for all i by monotonicity of the test levels and candidacy thresholds. This concludes the proof of the first inequality.

The second inequality uses a similar idea of hallucinating tests with identical finish times, only now the p-values that these tests result in are:

$$\widetilde{P}_i = \begin{cases} 0 & \text{if } i = t, \\ P_i & \text{if } i \neq t, \end{cases}$$

where P_i are the p-values in the original sequence. In a similar fashion, the following observations hold:

1. $\widetilde{E}_j = E_j$ implies $\alpha_i = \widetilde{\alpha}_i$ for all $i \le E_t$.
2. $\widetilde{\mathcal{R}}_i = \mathcal{R}_i$ and $\widetilde{\mathcal{C}}_i = \mathcal{C}_i$ for all $i < E_t$, since the p-values from the finished tests and the respective test levels are the same in the original and hallucinated setting.
3. $\widetilde{\mathcal{R}}_{E_t} \supseteq \mathcal{R}_{E_t}$ and $\widetilde{\mathcal{C}}_{E_t} \supseteq \mathcal{C}_{E_t}$, and hence $\widetilde{\mathcal{R}}_i \supseteq \mathcal{R}_i$ also for all $i > E_t$, due to monotonicity of the test levels α_i.

Then, on the event $\{P_t \le \alpha_t\}$, we have $\mathcal{R}_{E_t} = \widetilde{\mathcal{R}}_{E_t}$ and $\mathcal{C}_{E_t} = \widetilde{\mathcal{C}}_{E_t}$, and hence also $\mathcal{R}_{1:M} = \widetilde{\mathcal{R}}^{t\to 1}_{1:M}$ and $\mathcal{C}_{1:M} = \widetilde{\mathcal{C}}^{t\to 1}_{1:M}$. From this we conclude that:

$$\frac{\mathbf{1}\{P_t \le \alpha_t\}}{g(|\mathcal{R}|_{1:M})} = \frac{\mathbf{1}\{P_t \le \alpha_t\}}{g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})}.$$

As in the first part of the proof, we use the fact that the null p-values are mutually independent and independent of the non-nulls, which allows us to conclude that $\widetilde{\mathcal{R}}^{t\to 1}_{1:M}$ is independent of P_t conditioned on $\mathcal{F}^{\mathcal{X}^{E_t}}_{\text{async}}$. This observation results in the following:

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{1}\{P_t \le \alpha_t\}}{g(|\mathcal{R}|_{1:M})} \;\middle|\; \mathcal{F}^{\mathcal{X}^{E_t}}_{\text{async}}\right] = \mathbb{E}\left[\frac{\mathbf{1}\{P_t \le \alpha_t\}}{g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})} \;\middle|\; \mathcal{F}^{\mathcal{X}^{E_t}}_{\text{async}}\right] &\le \mathbb{E}\left[\frac{\alpha_t}{g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})} \;\middle|\; \mathcal{F}^{\mathcal{X}^{E_t}}_{\text{async}}\right] \\ &\le \mathbb{E}\left[\frac{\alpha_t}{g(|\mathcal{R}|_{1:M})} \;\middle|\; \mathcal{F}^{\mathcal{X}^{E_t}}_{\text{async}}\right],\end{aligned}$$

where the first inequality follows by taking an expectation only with respect to P_t by invoking the asynchronous super-uniformity property (6), and the second inequality follows because $g(|\mathcal{R}|_{1:M}) \le g(|\widetilde{\mathcal{R}}|^{t\to 1}_{1:M})$ since $|\mathcal{R}_i| \le |\widetilde{\mathcal{R}}_i|$ for all i by monotonicity of the test levels. This concludes the proof of the lemma.

A.5 Proof of Theorem 5

LORD$_{\text{async}}$. Fix a time step t. First we show the claim for LORD$_{\text{async}}$, so suppose that

$$\widehat{\text{FDP}}_{\text{LORD}_{\text{async}}}(t) := \frac{\sum_{j\le t}\alpha_j}{\sum_{j\le t}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\} \vee 1} \le \alpha.$$

Then:

$$\begin{aligned}\text{FDR}(t) := \mathbb{E}\left[\frac{|\mathcal{V}(t)|}{|\mathcal{R}(t)| \vee 1}\right] &= \mathbb{E}\left[\frac{\sum_{j\le t, j\in\mathcal{H}^0}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\}}{\sum_{j\le t}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\} \vee 1}\right] \\ &\le \sum_{i\le t, i\in\mathcal{H}^0}\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le \alpha_i\}}{\sum_{j\le t}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\} \vee 1}\right],\end{aligned}$$

where the second equality follows by definition of $\mathcal{V}(t)$ and $\mathbb{R}(t)$, and the inequality drops the condition $E_i \le t$ from the numerator and applies linearity of expectation. Now we can apply Lemma 4 with $g(|\mathcal{R}|_{1:t}) = (\sum_{i=1}^{t}|\mathcal{R}_i|) \vee 1$, together with iterated expectations, to obtain:

$$\begin{aligned}\sum_{i\le t, i\in\mathcal{H}^0}\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le \alpha_i\}}{\sum_{j\le t}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\} \vee 1}\right] &\le \sum_{i\le t, i\in\mathcal{H}^0}\mathbb{E}\left[\frac{\alpha_i}{\sum_{j\le t}\mathbf{1}\{P_j \le \alpha_j, E_j \le t\} \vee 1}\right] \\ &\le \mathbb{E}\left[\widehat{\text{FDP}}_{\text{LORD}_{\text{async}}}(t)\right] \le \alpha,\end{aligned}$$

where the second inequality follows by dropping the condition $i \in \mathcal{H}^0$. This completes the proof for LORD$_{\text{async}}$.

SAFFRON$_{\text{async}}$. Now we move on to SAFFRON$_{\text{async}}$. Using the same steps as above, for any fixed time t, we can conclude the following inequality:

$$\text{FDR}(t) \leq \sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1}\right].$$

Here we additionally apply the other inequality of Lemma 4, with the same choice $g(|\mathcal{R}|_{1:t}) = (\sum_{i=1}^{t} |\mathcal{R}_i|) \vee 1$, again with iterated expectations:

$$\sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1}\right] \leq \sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_i \mathbf{1}\left\{P_i > \lambda_i\right\}}{(1-\lambda_i)(\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1)}\right].$$

Assuming that the inequality

$$\widehat{\text{FDP}}_{\text{SAFFRON}_{\text{async}}}(t) := \frac{\sum_{j \leq t} \frac{\alpha_j}{1-\lambda_j}(\mathbf{1}\left\{P_j > \lambda_j, E_j \leq t\right\} + \mathbf{1}\left\{E_j > t\right\})}{\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1} \leq \alpha$$

holds, it follows that:

$$\begin{aligned}
&\sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\alpha_i \mathbf{1}\left\{P_i > \lambda_i\right\}}{(1-\lambda_i)(\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1)}\right] \\
&\leq \mathbb{E}\left[\frac{\sum_{j \leq t} \frac{\alpha_j}{1-\lambda_j}(\mathbf{1}\left\{P_j > \lambda_j, E_j \leq t\right\} + \mathbf{1}\left\{E_j > t\right\})}{\sum_{j \leq t} \mathbf{1}\left\{P_j \leq \alpha_j, E_j \leq t\right\} \vee 1}\right] = \mathbb{E}\left[\widehat{\text{FDP}}_{\text{SAFFRON}_{\text{async}}}(t)\right] \leq \alpha,
\end{aligned}$$

where the first inequality follows by dropping the conditions $j \in \mathcal{H}^0$ and $\{P_j > \lambda_j\}$ for some rounds. The second inequality follows by assumption, hence proving the theorem.

A.6 Proof of Theorem 6

For statement (a), we begin by noting that for any $t \in \mathbb{N}$:

$$\text{FDR}(t) = \mathbb{E}\left[\frac{\sum_{i \leq t, i \in \mathcal{H}^0} \mathbf{1}\left\{P_i \leq \alpha_i\right\}}{|\mathcal{R}(t)| \vee 1}\right] \leq \sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_i\right\}}{|\mathcal{R}(i-1)| \vee 1}\right] = \sum_{i \leq t, i \in \mathcal{H}^0} \gamma_i \alpha \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_i\right\}}{\alpha_i}\right],$$

where the first equality follows by definition of FDR, the sole inequality follows because the number of rejections can only increase with time, and the second equality follows by definition of the LOND rule for α_i. Lemma 1 from Ramdas et al. (2019) now asserts that the term in the expectation is bounded by one under PRDS. Hence, by also noting that $\sum_{i \leq t} \gamma_i \leq 1$ we immediately deduce statement (a).

For statement (b), we follow almost the same sequence of steps to note that:

$$\begin{aligned}
\text{FDR}(t) &= \mathbb{E}\left[\frac{\sum_{i \leq t, i \in \mathcal{H}^0} \mathbf{1}\left\{P_i \leq \widetilde{\alpha}_i\right\}}{|\mathcal{R}(t)| \vee 1}\right] \leq \sum_{i \leq t, i \in \mathcal{H}^0} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \widetilde{\alpha}_i\right\}}{|\mathcal{R}(i-1)| \vee 1}\right] \\
&= \sum_{i \leq t, i \in \mathcal{H}^0} \gamma_i \alpha \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \gamma_i \alpha \beta_i (|\mathcal{R}(i-1)| \vee 1)\right\}}{\gamma_i \alpha (|\mathcal{R}(i-1)| \vee 1)}\right].
\end{aligned}$$

We now apply Lemma 1 from Ramdas et al. (2019) with $c = \gamma_i \alpha$ and $f(P) = |\mathcal{R}(i-1)| \vee 1$ to again assert that the term in the expectation is bounded by one under arbitrary dependence, hence establishing statement (b).

Appendix B. Different instantiations of LORD* and SAFFRON*

Here we give explicit statements of different instances of LORD* and SAFFRON* described in Section 3, Section 4 and Section 5. All of the following algorithms are special instances of Algorithms 1-4, given in Section 2.

First we state $\text{LORD}_{\text{async}}$ and $\text{SAFFRON}_{\text{async}}$ explicitly, by taking $\mathcal{X}^t = \mathcal{X}^t_{\text{async}}$ in the statement of LORD* and SAFFRON*. Algorithm 5 and Algorithm 6 state the LORD++ and LOND versions of $\text{LORD}_{\text{async}}$, Algorithm 7 states $\text{SAFFRON}_{\text{async}}$ for constant candidacy thresholds, i.e. $\{\lambda_j\} \equiv \lambda$, and Algorithm 8 states asynchronous alpha-investing, i.e. $\text{SAFFRON}_{\text{async}}$ when $\lambda_j = \alpha_j$. Recall the definition of r_k, which in this setting takes the form:

$$r_k = \min\{i \in [t] : \sum_{j=1}^{i} R_j \mathbf{1}\{E_j \le i\} \ge k\}.$$

Algorithm 5 The asynchronous LORD++ algorithm as a version of $\text{LORD}_{\text{async}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \le \alpha$

$\alpha_1 = \gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

  start t-th test with level α_t

  $\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-r_1}(\alpha - W_0) + \left(\sum_{j \ge 2} \gamma_{t+1-r_j}\right) \alpha$

end

Algorithm 6 The asynchronous LOND algorithm as a version of $\text{LORD}_{\text{async}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$

$W_0 = \alpha$

$\alpha_1 = \gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

  start t-th test with level α_t

  $\alpha_{t+1} = \alpha \gamma_{t+1} \left((\sum_{j=1}^{t} \mathbf{1}\{P_j \le \alpha_j, E_j \le t\}) \vee 1 \right)$

end

Algorithm 7 The $\text{SAFFRON}_{\text{async}}$ algorithm for constant λ

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, candidate threshold $\lambda \in (0, 1)$, initial wealth $W_0 \le \alpha$

$\alpha_1 = (1 - \lambda)\gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

  start t-th test with level α_t

  $\alpha_{t+1} = \min\left\{\lambda, (1-\lambda)\left(W_0 \gamma_{t+1-C^{\#}_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-C^{\#}_{1+}} + \sum_{j \ge 2} \alpha \gamma_{t+1-r_j-C^{\#}_{j+}}\right)\right\}$,

  where $C^{\#}_{j+} = \sum_{i=r_j+1}^{t} |\mathcal{C}_i|$

end

Algorithm 8 The asynchronous alpha-investing algorithm as a special case of $\text{SAFFRON}_{\text{async}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \le \alpha$

$s_1 = \gamma_1 W_0$

$\alpha_1 = s_1/(1 + s_1)$

for $t = 1, 2, \ldots$ **do**

  start t-th test with level α_t

  $s_{t+1} = W_0 \gamma_{t+1-R^{\#}_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-R^{\#}_{1+}} + \sum_{j \ge 2} \alpha \gamma_{t+1-r_j-R^{\#}_{j+}}$, where $R^{\#}_{j+} = \sum_{i=r_j+1}^{t} |\mathcal{R}_i|$

  $\alpha_{t+1} = s_{t+1}/(1 + s_{t+1})$

end

Below we give explicit statements of LORD_{dep} and $\text{SAFFRON}_{\text{dep}}$ as special cases of LORD* and SAFFRON*. Algorithm 9 and Algorithm 10 state LORD++ and LOND under local dependence, both as instances of LORD_{dep}. Algorithm 11 states $\text{SAFFRON}_{\text{dep}}$ for the constant sequence $\{\lambda_j\} \equiv \lambda$, and Algorithm 12 states alpha-investing under local dependence, which is a particular instance of $\text{SAFFRON}_{\text{dep}}$ obtained by taking $\lambda_j = \alpha_j$. The definition of r_k under local dependence simplify to:

$$r_k = \min\{i \in [t] : \sum_{j=1}^{i-L_{i+1}} R_j \geq k\}.$$

Algorithm 9 The LORD++ algorithm under local dependence as a version of LORD_{dep}

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \leq \alpha$
$\alpha_1 = \gamma_1 W_0$
for $t = 1, 2, \ldots$ **do**
  run t-th test with level α_t
  $\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-r_1}(\alpha - W_0) + \left(\sum_{j=2}^{\infty} \gamma_{t+1-r_j}\right) \alpha$
end

Algorithm 10 The LOND algorithm under local dependence as a version of LORD_{dep}

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$
$W_0 = \alpha$
$\alpha_1 = \gamma_1 W_0$
for $t = 1, 2, \ldots$ **do**
  run t-th test with level α_t
  $\alpha_{t+1} = \alpha \gamma_{t+1} \left((\sum_{i=1}^{t-L_{t+1}} R_i) \vee 1\right)$
end

Algorithm 11 The $\text{SAFFRON}_{\text{dep}}$ algorithm for constant λ

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, candidate threshold $\lambda \in (0, 1)$, initial wealth $W_0 \leq \alpha$
$\alpha_1 = \gamma_1 W_0$
for $t = 1, 2, \ldots$ **do**
  run t-th test with level α_t
  $\alpha_{t+1} = \min\left\{\lambda, (1-\lambda)\left(W_0 \gamma_{t+1-C_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-C_{1+}} + \alpha(\sum_{j\geq 2} \gamma_{t+1-r_j-C_{j+}})\right)\right\}$,
  where $C_{j+} = \sum_{i=r_j+1}^{t-L_{t+1}} C_i$
end

Algorithm 12 The alpha-investing algorithm under local dependence as a special case of $\text{SAFFRON}_{\text{dep}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_0 \leq \alpha$
$s_1 = \gamma_1 W_0$
$\alpha_1 = s_1/(1 + s_1)$
for $t = 1, 2, \ldots$ **do**
  run t-th test with level α_t
  $s_{t+1} = W_0 \gamma_{t+1-R_{0+}} + (\alpha - W_0)\gamma_{t+1-r_1-R_{1+}} + \sum_{j\geq 2} \alpha \gamma_{t+1-r_j-R_{j+}}$, where $R_{j+} = \sum_{i=r_j+1}^{t-L_{t+1}} R_i$
  $\alpha_{t+1} = s_{t+1}/(1 + s_{t+1})$
end

Algorithms 13 and 14 describe the mini-batch versions of LORD++ and LOND respectively, both as cases of $\text{LORD}_{\text{mini}}$. Algorithm 15 is a variant of $\text{SAFFRON}_{\text{mini}}$ with λ_j chosen constant and equal to some

$\lambda \in (0, 1)$, and Algorithm 16 is the alpha-investing version of SAFFRON$_{\text{mini}}$, in which $\lambda_j = \alpha_j$. In this setting, the definition of r_k is slightly tweaked in order to satisfy the convention of double indexing; r_k refers to the *batch* in which the k-th non-conflicting rejection occurs:

$$r_k := \min\{i \in [b-1] : \sum_{j=1}^{i} |\mathcal{R}_j| \geq k\}.$$

Algorithm 13 The mini-batch LORD++ algorithm as a version of LORD$_{\text{mini}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_{1,0} \leq \alpha$
$\alpha_{1,1} = \gamma_1 W_{1,0}$
for $b = 1, 2, \ldots$ **do**
  if $b > 1$ **then**
    $W_{b,0} = W_{b-1,n} + \alpha|\mathcal{R}_{b-1}| - W_{1,0}\mathbf{1}\{r_1 = b-1\}$
  end
  for $t = 1, 2, \ldots, n_b$ **do**
    start t-th test in the b-th batch with level $\alpha_{b,t}$
    $\alpha_{b,t+1} = \gamma_{\sum_{i=1}^{b-1} n_i + t + 1} W_{1,0} + \gamma_{\sum_{i=1}^{b-1} n_i + t + 1 - \sum_{i=1}^{r_1} n_i}(\alpha - W_{1,0}) + \left(\sum_{j=2}^{\infty} \gamma_{\sum_{i=1}^{b-1} n_i + t + 1 - \sum_{i=1}^{r_j} n_i}\right)\alpha$
  end
end

Algorithm 14 The mini-batch LOND algorithm as a version of LORD$_{\text{mini}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$
$\alpha_{1,1} = \gamma_1 \alpha$
for $b = 1, 2, \ldots$ **do**
  for $t = 1, 2, \ldots, n_b$ **do**
    start t-th test in the b-th batch with level $\alpha_{b,t}$
    $\alpha_{b,t+1} = \alpha\gamma_{\sum_{i=1}^{b-1} n_i + t + 1}\left((\sum_{j=1}^{b-1} |\mathcal{R}_j|) \vee 1\right)$
  end
end

Algorithm 15 The SAFFRON$_{\text{mini}}$ algorithm for constant λ

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_{1,0} \leq \alpha$, constant λ
$\alpha_{1,1} = (1-\lambda)\gamma_1 W_{1,0}$
for $b = 1, 2, \ldots$ **do**
  for $t = 1, 2, \ldots, n$ **do**
    start t-th test in the b-th batch with level $\alpha_{b,t}$
    $\alpha_{b,t+1} = (1-\lambda)(\gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_0^+| + t + 1} W_{1,0} + \gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_1^+| + t + 1 - \sum_{i=1}^{r_1} n_i}(\alpha - W_{1,0}) + \left(\sum_{j=2}^{\infty} \gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_j^+| + t + 1 - \sum_{i=1}^{r_1} n_i}\right)\alpha)$, where $|\mathcal{C}_j^+| = \sum_{j=r_j+1}^{b-1} |\mathcal{C}_j|$
  end
end

Algorithm 16 The mini-batch alpha-investing as a special case of SAFFRON$_{\text{mini}}$

input: FDR level α, non-negative non-increasing sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_j \gamma_j = 1$, initial wealth $W_{1,0} \leq \alpha$

$s_{1,1} = \gamma_1 W_{1,0}$

$\alpha_{1,1} = s_{1,1}/(1 + s_{1,1})$

for $b = 1, 2, \ldots$ **do**

  for $t = 1, 2, \ldots, n$ **do**

    start t-th test in the b-th batch with level $\alpha_{b,t}$

    $s_{b,t+1} = \gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_0^+| + t + 1} W_{1,0} + \gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_1^+| + t + 1 - \sum_{i=1}^{r_1} n_i} (\alpha - W_{1,0}) + \left(\sum_{j=2}^{\infty} \gamma_{\sum_{i=1}^{b-1} n_i - |\mathcal{C}_j^+| + t + 1 - \sum_{i=1}^{r_j} n_i} \right) \alpha$, where $|\mathcal{C}_j^+| = \sum_{j=r_j+1}^{b-1} |\mathcal{C}_j|$

    $\alpha_{b,t+1} = s_{b,t+1}/(1 + s_{b,t+1})$

  end

end

Appendix C. Experiments on real data with local dependence

We perform an additional case study on a high-throughput phenotypic data set from the International Mouse Phenotyping Consortium (IMPC) data repository. This database documents gene knockout experiments on mice and annotates every protein coding gene by exploring the impact of the gene knockout on the resulting phenotype. This is an example of a continuously growing data set, as the family of hypotheses and the new knockouts grow with time. Karp et al. (2017) tested the role of genotype and the role of sex as a modifier of genotype effect. In this section, we focus on the p-values resulting from the analysis of genotype effects. This set of p-values exhibits local dependence—the same set of mice is used to test multiple hypotheses adjacent on the time horizon, while p-values computed at sufficiently distant time points are statistically independent.

We use the subset of the database organized by Robertson et al. (2019), which is available at `https://zenodo.org/record` Hypotheses belonging to the same batch have the same experimental ID. For the sake of computational efficiency, we only analyze the hypotheses whose experimental ID is in the interval $[36700, 37000)$. This results in 4275 distinct hypotheses and their corresponding p-values, split into 172 batches of varying sizes. For different target FDR levels, we report the number of discoveries made by SAFFRON$_{\text{dep}}$, LORD$_{\text{dep}}$, and alpha-spending. Since we expect a small number of truly relevant genes, for SAFFRON$_{\text{dep}}$ we set $\lambda = 0.1$; all other parameters for all three algorithms are as in Section 7. We cannot evaluate the FDR and power due to lack of ground truth, but theoretically all three procedures control the FDR at the target level.

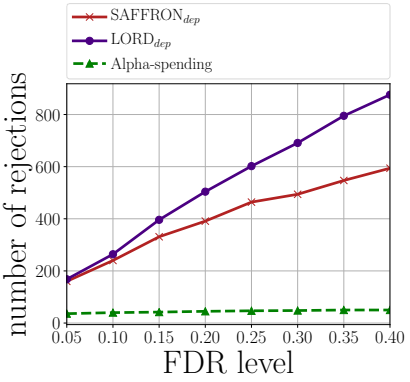

Figure 12: Number of rejections made by SAFFRON$_{\text{dep}}$, LORD$_{\text{dep}}$, and alpha-spending, for different target FDR levels.

Appendix D. Additional experiments with varying asynchrony

We illustrate the tradeoff between the power gain of optimized tests, which make use of the test level α_i at the beginning of the test, and the accompanying power loss when running such tests asynchronously. We do so by contrasting optimized sequential tests with their simpler counterparts which compute p-values without knowledge of α_i in a simulated A/B testing environment.

We test null hypotheses of the form $H_i : \mu^{(i)}_{\text{treatment}} - \mu^{(i)}_{\text{control}} = 0$, and for each test we collect samples from $N(\mu^{(i)}_{\text{control}}, 1)$ and $N(\mu^{(i)}_{\text{treatment}}, 1)$. We collect samples by uniformly sampling from the treatment and control distribution; at every sampling step j of test i we draw a single sample from both distributions, $X^{(i,j)}_{\text{treatment}} \sim N(\mu^{(i)}_{\text{treatment}}, 1)$, $X^{(i,j)}_{\text{control}} \sim N(\mu^{(i)}_{\text{control}}, 1)$, akin to allocating one of two matched patients to each of those arms. We contrast two approaches: in one, we fix the number of per-test draws n from each distribution, and simply calculate a p-value by computing $P_i = 1 - \Phi(\sqrt{n/2}(\hat{\mu}^{(i)}_{\text{treatment}} - \hat{\mu}^{(i)}_{\text{control}}))$, where $\hat{\mu}^{(i)}_{\text{treatment}} = \frac{1}{n}\sum_{j=1}^{n} X^{(i,j)}_{\text{treatment}}$ and $\hat{\mu}^{(i)}_{\text{control}} = \frac{1}{n}\sum_{j=1}^{n} X^{(i,j)}_{\text{control}}$. This approach does not require an asynchronous correction. In the other approach, we use the test level α_i and the law of the iterated logarithm to *adaptively* stop sample collection once a rejection is admissible; as in the first approach, we collect at most n samples from each distribution, but we stop and make a rejection after $k \leq n$ draws if

$$\sum_{j=1}^{k} \left(X^{(i,j)}_{\text{treatment}} - X^{(i,j)}_{\text{control}} \right) > \sqrt{k(4\log(1/(\alpha_i \wedge 0.1)) + 12\log(\log(1/(\alpha_i \wedge 0.1))) + 6\log(\log(ek)))}.$$

Validity of this stopping rule follows from Theorem 8 in Kaufmann et al. (2016). For other stopping rules that can be employed in nonparametric situations, see Howard et al. (2021).

Rather than fixing the total number of tests, we fix the total sample budget B. In the simpler strategy, this corresponds to fixing the total number of tests one can perform, namely $\lfloor B/(2n) \rfloor$. In the adaptive approach, the number of tests is lower bounded by $\lfloor B/(2n) \rfloor$, but can be higher if rejections are made with fewer than n samples drawn per distribution.

For each test, with equal probability we let $\mu^{(i)}_{\text{treatment}} - \mu^{(i)}_{\text{control}} = 0$, in which case the null hypothesis is true, or we draw the margin from an exponential distribution, $\mu^{(i)}_{\text{treatment}} - \mu^{(i)}_{\text{control}} \sim \text{Exp}(1)$, in which case the null is false. We set $\alpha = 0.05$ and average all results over 200 trials. In each trial, all considered algorithms observe the same sequence of average treatment effects, $\{\mu^{(i)}_{\text{treatment}} - \mu^{(i)}_{\text{control}}\}$, but draw independent Gaussian samples. In Figure 13, we plot the number of true discoveries made by LORD against different values of the sample budget B for several different values of n (varying n changes the false negative rate). The adaptive stopping strategy with a low level of asynchrony generally outperforms the naive approach; however, if the level of asynchrony is high (relative to other problem parameters), the naive approach dominates, as expected.

Appendix E. Positive regression dependency on a subset (PRDS)

We briefly review the definition of positive regression dependency on a subset (PRDS).

Definition 7 *Let $\mathcal{D} \subseteq [0,1]^n$ be any non-decreasing set, meaning that $x \in \mathcal{D}$ implies $y \in \mathcal{D}$, for all y such that $y_i \geq x_i$ for all $i \in [n]$. We say that a vector of p-values $P = (P_1, \ldots, P_n)$ satisfies positive dependence (PRDS) if for any null index $i \in \mathcal{H}^0$ and arbitrary non-decreasing $\mathcal{D} \subseteq [0,1]^n$, the function $t \to Pr\{P \in \mathcal{D} \mid P_i \leq t\}$ is non-decreasing over $t \in (0,1]$.*

Clearly, independent p-values satisfy PRDS. Another important example is given for Gaussian observations. Suppose $Z = (Z_1, \ldots, Z_n)$ is a multivariate Gaussian with covariance matrix Σ, and let $P = (\Phi(Z_1,), \ldots, \Phi(Z_n))$ be a vector of p-values, where Φ is the standard Gaussian CDF. Then, P satisfies PRDS if and only if, for all $i \in \mathcal{H}^0$ and $j \in [n]$, $\Sigma_{ij} \geq 0$.

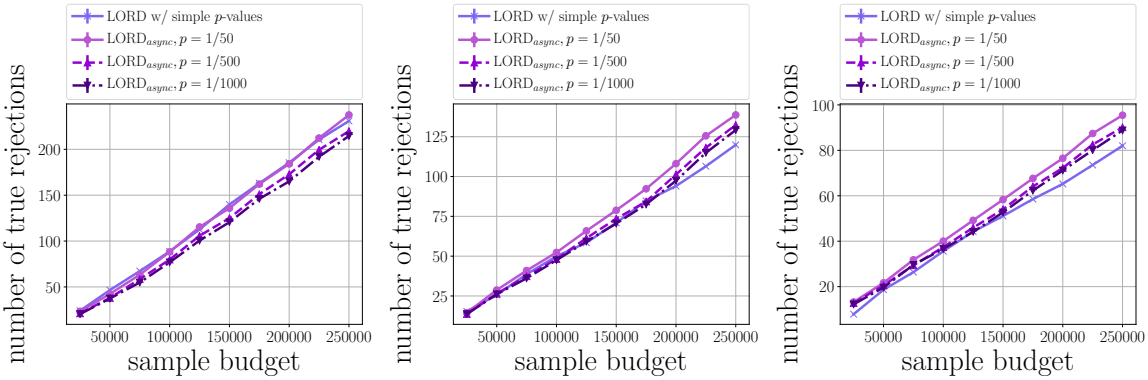

Figure 13: Number of true discoveries made by LORD with simple p-values and $\text{LORD}_{\text{async}}$ with a sequential strategy, against the sample budget B. The values of n are 400 (left), 800 (middle), 1200 (right). The duration of asynchronous tests is sampled as in Section 7.

Appendix F. Examining the difference between mFDR and FDR

In Section 7, we plotted strict FDR estimates, obtained by averaging the false discovery proportion over 200 independent trials; on the other hand, the main guarantees of this paper apply to mFDR control. For this reason, here we provide the plot of both mFDR and FDR estimates, for all experiments in Section 7. We estimate mFDR by computing the ratio of the average number of false discoveries and the average total number of discoveries.

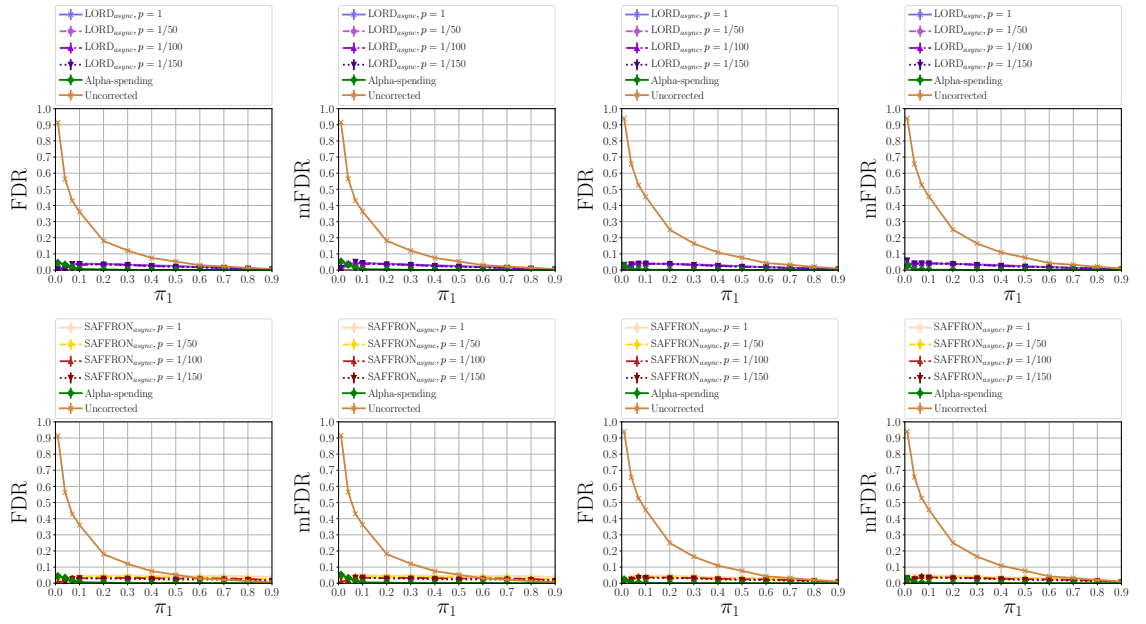

Figure 14: The left plots reproduce FDR from Figure 5 and Figure 6, while the right plots show mFDR for the same experiments.

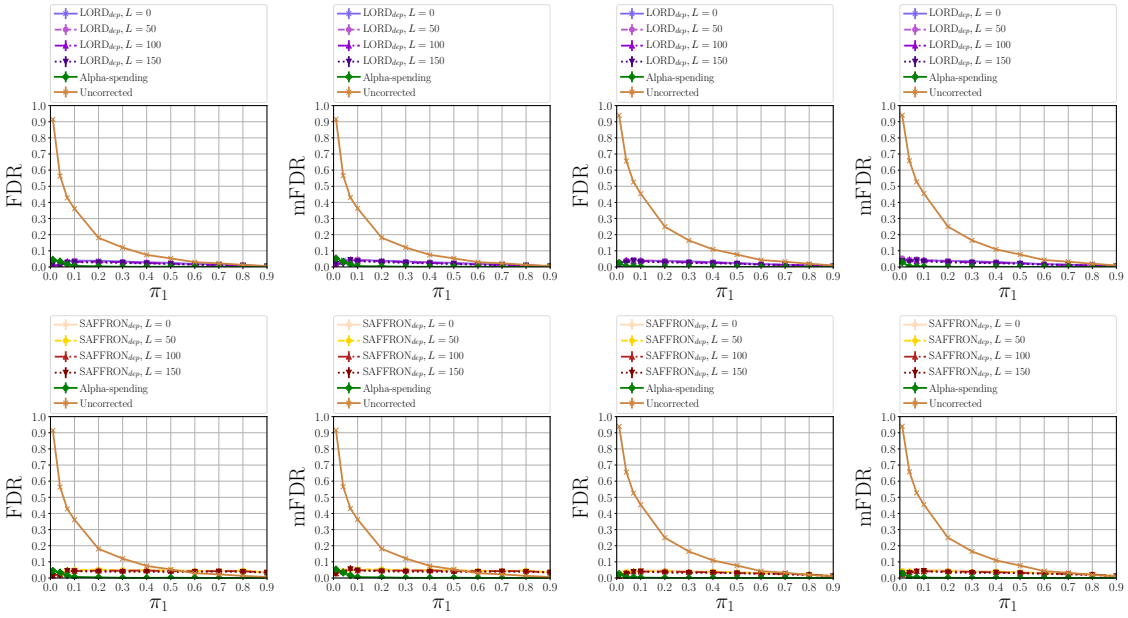

Figure 15: The left plots reproduce FDR from Figure 7 and Figure 8, while the right plots show mFDR for the same experiments.

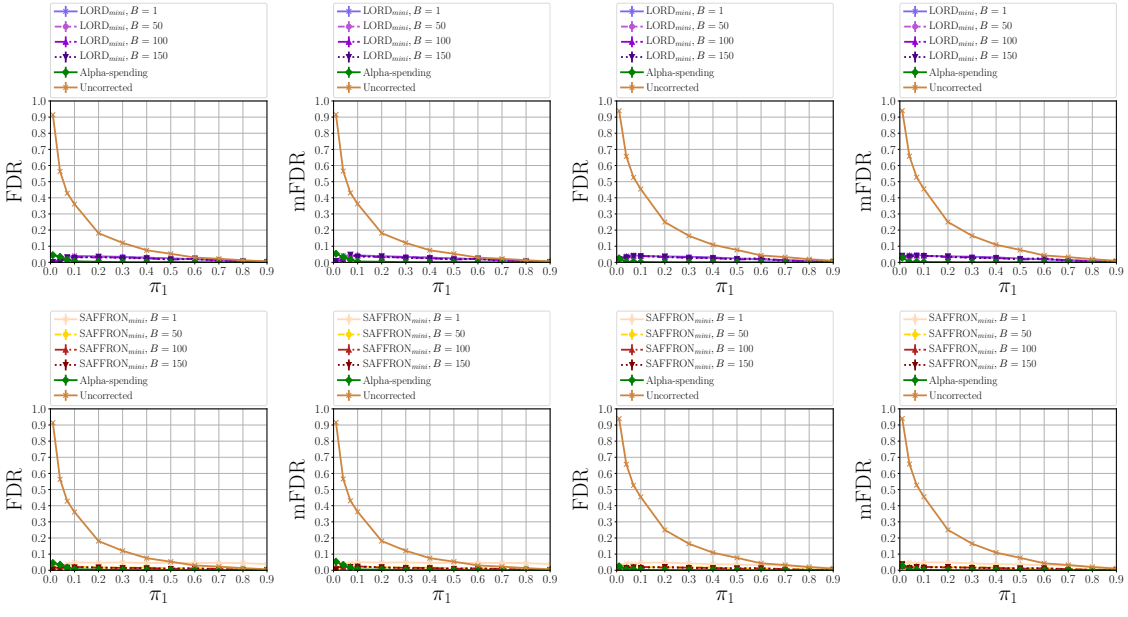

Figure 16: The left plots reproduce FDR from Figure 9 and Figure 10, while the right plots show mFDR for the same experiments.

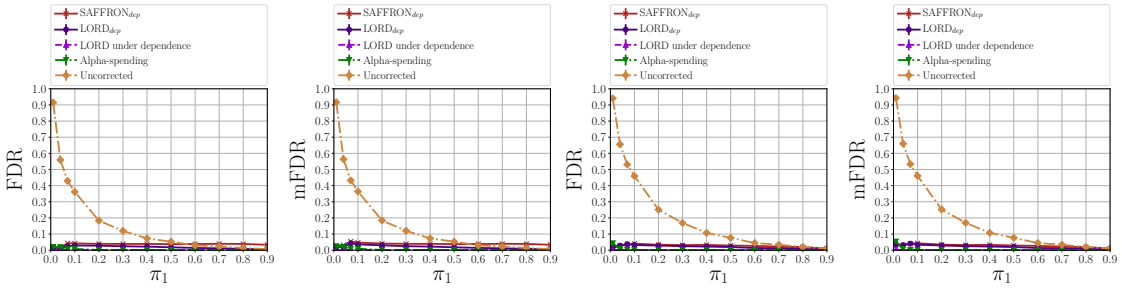

Figure 17: The left plots reproduce FDR from Figure 11, while the right plots show mFDR for the same experiments.

References

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

Arthur E Albert. The sequential design of experiments for infinitely many states of nature. *The Annals of Mathematical Statistics*, pages 774–799, 1961.

Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016.

Jay Bartroff and Tze Leung Lai. Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Sequential Analysis*, 27(3):254–276, 2008.

Jay Bartroff and Jinlin Song. Sequential tests of multiple hypotheses controlling type I and II familywise error rates. *Journal of statistical planning and inference*, 153:100–114, 2014.

Jay Bartroff, Tze Leung Lai, and Mei-Chiung Shih. *Sequential experimentation in clinical trials: design and analysis*, volume 298. Springer Science & Business Media, 2012.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

Yoav Benjamini and Yosef Hochberg. 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.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

Avrim Blum and Moritz Hardt. The Ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1006–1014, 2015.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 117–126. ACM, 2015b.

Dean P Foster and Robert A Stine. 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.

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

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(2):423–444, 2016.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics (to appear)*, 2021.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Kevin Jamieson and Lalit Jain. A bandit approach to multiple testing with false discovery control. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3664–3674, 2018.

Adel Javanmard and Andrea Montanari. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

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

Natasha A Karp, Jeremy Mason, Arthur L Beaudet, Yoav Benjamini, Lynette Bower, Robert E Braun, Steve DM Brown, Elissa J Chesler, Mary E Dickinson, Ann M Flenniken, et al. Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nature Communications*, 8(1):1–12, 2017.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Lihua Lei and William Fithian. Power of ordered hypothesis testing. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2924–2932, 2016.

Ang Li and Rina Foygel Barber. Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017.

Gavin Lynch, Wenge Guo, Sanat K Sarkar, and Helmut Finner. The control of the false discovery rate in fixed sequence multiple testing. *Electronic Journal of Statistics*, 11(2):4649–4673, 2017.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6): 2703–2738, 2013.

Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5655–5664, 2017.

Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. SAFFRON: an adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4286–4294, 2018.

Aaditya K Ramdas, Rina F Barber, Martin J Wainwright, Michael I Jordan, et al. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47(5):2790–2821, 2019.

David S Robertson and James Wason. Online control of the false discovery rate in biomedical research. *arXiv preprint arXiv:1809.07292v1*, 2018.

David S Robertson, Jan Wildenhain, Adel Javanmard, and Natasha A Karp. onlinefdr: An R package to control the false discovery rate for growing data repositories. *Bioinformatics*, 35(20):4196–4199, 2019.

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

John D Storey, Jonathan E Taylor, and David Siegmund. 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.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 17–26, 2010.

Edwin JCG van den Oord. Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(5):637–644, 2008.

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

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236, 2015.

Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. A framework for multi-A (rmed)/B (andit) testing with online FDR control. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5957–5966, 2017.