Contextual Online False Discovery Rate Control

Shiyun Chen*   Shiva Kasiviswanathan†

Abstract

Multiple hypothesis testing, a situation when we wish to consider many hypotheses, is a core problem in statistical inference that arises in almost every scientific field. In this setting, controlling the false discovery rate (FDR), which is the expected proportion of type I error, is an important challenge for making meaningful inferences. In this paper, we consider the problem of controlling FDR in an online manner. Concretely, we consider an ordered, possibly infinite, sequence of hypotheses, arriving one at each timestep, and for each hypothesis we observe a p-value along with a set of features specific to that hypothesis. The decision whether or not to reject the current hypothesis must be made immediately at each timestep, before the next hypothesis is observed. The model of multi-dimensional feature set provides a very general way of leveraging the auxiliary information in the data which helps in maximizing the number of discoveries.

We propose a new class of powerful online testing procedures, where the rejections thresholds (significance levels) are learnt sequentially by incorporating contextual information and previous results. We prove that any rule in this class controls online FDR under some standard assumptions. We then focus on a subclass of these procedures, based on weighting significance levels, to derive a practical algorithm that learns a parametric weight function in an online fashion to gain more discoveries. We also theoretically prove, in a stylized setting, that our proposed procedures would lead to an increase in the achieved statistical power over a popular online testing procedure proposed by Javanmard & Montanari (2018). Finally, we demonstrate the favorable performance of our procedure, by comparing it to state-of-the-art online multiple testing procedures, on both synthetic data and real data generated from different applications.

---

*Department of Mathematics, University of California, San Diego. shc176@ucsd.edu. Part of the work done with the author was an intern at Amazon.

†Amazon, Sunnyvale, USA. kasivisw@gmail.com.

Contents

1 Introduction

Multiple hypotheses testing - controlling overall error rates when performing simultaneous hypothesis tests - is a well-established area in statistics with applications in a variety of scientific disciplines (Dickhaus, 2014; Dudoit & van der Laan, 2007; Roquain, 2011). This problem has become even more important with modern data science, where standard data pipelines involve performing a large number of hypotheses tests on complex datasets, e.g., does this change to my webpage improve my click-through rate, is this ad effective for this population of users, or is this gene mutation associated with certain trait?

In hypothesis testing problems, each hypothesis is summarized to one p-value, which we use to decide whether to reject the null hypothesis, i.e., claim it as a non-null. Typically, a hypothesis is rejected if p-value is below some significance level. The rejected hypotheses are called *discoveries*, and the subset of these that were truly null but mistakenly rejected are called *false discoveries*. The *false discovery rate* (FDR) namely, the expected fraction of discoveries that are false positives is the criterion of choice for statistical inference in multiple hypothesis testing problems. The traditional multiple testing research has focused on the offline setting, which means we have an entire batch of hypotheses and the corresponding p-values. In their seminal work, Benjamini & Hochberg (1995) developed a procedure (called *BH procedure*) to control FDR below a preassigned level for this offline setting. However, the fact that offline FDR control techniques require aggregating p-values from all the tests and processing them jointly, makes it impossible to utilize them for a number of applications which are best modeled as an *online hypothesis testing* problem (Foster & Stine, 2008) (a more formal definition will be provided later). In this scenario, we assume that an infinite sequence of hypotheses arrive sequentially in a stream, and decisions are made only based on previous results before next hypothesis arrives. In other words, the decisions have to be made without access to the number of hypotheses in the stream or the future p-values, but solely based on the previous decisions. For example, in marketing research a sequence of A/B tests can be carried out in an online fashion, or in a pharmaceutical drug test a sequence of clinical trial are conducted over time, or with publicly available datasets where new hypotheses are tested in an on-going fashion by different researchers.

Foster & Stine (2008) designed the first online alpha-investing procedures that use and earn alpha-wealth in order to control a modified definition of FDR (referred to as *mFDR*), which was later extended to a class of *generalized alpha-investing* (GAI) rules by Aharoni & Rosset (2014). Javanmard & Montanari (2015, 2018) showed that monotone GAI rules, appropriately parametrized, can control the FDR for independent p-values as opposed to the modified FDR controlled in (Aharoni & Rosset, 2014; Foster & Stine, 2008). Within this class, of special note here is an procedure (testing rule) called *LORD* that the authors noted performs consistently well in practice. Ramdas et al. (2017b) demonstrated a modification to the GAI class (called GAI++) that improved the earlier GAI algorithms (uniformly), and the improved LORD++ method arguably represents the current state-of-the-art in online multiple hypothesis testing. Very recently, (Ramdas et al., 2018) empirically demonstrated that using adaptiveness, some further improvements in the power over LORD++ can be obtained. We survey these online testing procedures in more detail in Section 2.

All these above online testing procedures take p-values as input and decide at each time based on previous decision outcomes. However, these procedures ignore additional information that is often available in modern applications. In addition to the p-value P_i, each hypothesis H_i could also have a feature vector X_i lying in some space $\mathcal{X} \subseteq \mathbb{R}^d$. The feature vector encode auxiliary[1] information related to the tested hypothesis, and is also often referred as *context*. The feature vector X_i only carry indirect information about the likelihood of the hypothesis H_i to be false but the relationship is not fully known ahead of time. For example, when conducting an A/B test for a logo size change in a website, contextual information such as text, layouts, images and colors in this specific page can be useful in making a more informative decision. Similarly another example arises when testing whether a mutation is correlated with the trait, here contextual

---

[1]Also sometimes referred to as *prior* or *side* information.

information about both the mutation and the trait such as its location, epigenetic status, etc., could provide valuable information that can increase the power of these tests.

While the problem of using auxiliary information in testing has been considered in offline setting (Genovese et al., 2006; Ignatiadis et al., 2016; Lei & Fithian, 2018; Li & Barber, 2016b; Ramdas et al., 2017a; Xia et al., 2017), in this paper we focus on the more natural online setting, where unlike the batch setting p-values and contextual features are not available at the onset, and a decision about a hypothesis should be made when it is presented. To the best of our knowledge, this generalization of the online testing problem has not been considered before. Our main contributions in this paper are as follows.

(1) Incorporating Contextual Information. Building on GAI/GAI++ rules (Aharoni & Rosset, 2014; Ramdas et al., 2017b), we propose a new broad class of powerful online testing rules, that we refer to as *contextual generalized alpha-investing* (CGAI) rules. This new class of rules incorporates the available contextual features in the testing process, and we prove that any monotone rule from this class can control online FDR under some standard assumption on p-values. More formally, we assume each hypothesis H is characterized by a tuple (P, X) where $P \in (0, 1)$ is the p-value, X is the contextual feature vector. We use a very general model for these X's, as a vector coming from some generic space $\mathcal{X} \subseteq \mathbb{R}^d$. We consider a sequence of hypotheses $(H_1, H_2, \dots)$ that arrive sequentially in a stream one at each timestep $t = 1, 2, \dots$, with corresponding $((P_1, X_1), (P_2, X_2), \dots)$, our testing rule generates a sequence of significance levels $(\alpha_1, \alpha_2, \dots)$ at each timestep based on previous decisions and contextual information seen so far. Our test for each hypothesis H_t takes the form $\mathbb{1}\{P_t \leq \alpha_t\}$, and under mutual independence of the p-values and independence between the p-values P_t's and the contextual features X_t's for null hypotheses, we show that any monotone rule from this class ensures that FDR remains below a preassigned level at any time. In the proof the sigma-field formalizes "what we know" at time t, and we define a filtration via the sigma-fields of both previous decisions and the contextual features. We also show that we can have mFDR control under a weaker assumption on p-values.

(2) Context Weighting. While contextual generalized alpha-investing (CGAI) is a rich class of FDR control rules, we focus on a subclass of these rules for designing a practical online FDR control procedure and to compare the statistical power of various procedures. In particular, we focus on a subclass, that we refer to as *context-weighted generalized alpha-investing* (CwGAI) rules, where the contextual features are used for weighting the significance levels in an online setting. In particular, we take a parametric function $\omega(; \theta)$ for a set of parameters θ, and at timestep t use $\omega(X_t; \theta)$ as a real-valued weight on α_t generated through GAI rules, with the intuition that larger weights should reflect an increased willingness to reject the null. Since the parameter set θ is unknown, a natural idea here will be to learn it in an online fashion to maximize the number of empirical discoveries. This gives rise to a new class of online testing rules that incorporates the context weights through a learnt parametric function.

(3) Statistical Power Analysis. Having established that our procedures control the false discovery rate, we next turn to the question of what effect context weighting has on the statistical power in an online setting. Lots of factors, such as frequency of true hypotheses, or on their order, affect the power. Hence, to make a rigorous comparison, we consider a standard mixture model used in such analyses where each null hypothesis is false with a fixed (unknown) probability (Genovese et al., 2006; Javanmard & Montanari, 2018; Lei & Fithian, 2018), and focus on a slightly weaker variant of the popular online testing LORD procedure (Javanmard & Montanari, 2018). By considering a very general model of weighting where weights are random variables, and under the assumption that weights are positively associated with the null hypotheses being false, we derive a natural sufficient condition on the weights under which weighting improves the power in an online setting, while still guaranteeing FDR control. This is the first result that demonstrates the benefits of appropriate weighting in the online setting. Prior to this such results were only known in the offline setting (Genovese et al., 2006).

(4) A Practical Procedure. To design a practical online FDR control procedure with good performance, we model the parametric function $\omega(; \theta)$ as a neural network (multilayer perceptron), and train it in an online fashion to maximize the number of empirical discoveries. Our experimental evidence on a range of synthetic and real datasets show that our proposed procedure makes substantially more correct decisions compared to state-of-the-art online testing procedures.

We formally describe the online multiple testing problem setup and review the published literature in this area in Section 2. Here we start with a review of some relevant prior work in offline multiple testing.

1.1 Related Work in the Offline Setting

In the offline setting, where we have access to the entire batch of p-values at one time instant, a number of procedures have been proposed to take advantage of the available auxiliary information to increase the power of test (to make more true discoveries). As we note below, the modeling of auxiliary information varies.

Storey (2002) proposed an adaptive FDR-control procedure based on estimating the proportion of true nulls from data. Reweighting the p-values by applying priors was considered by (Benjamini & Hochberg, 1997; Dobriban, 2016; Dobriban et al., 2015; Genovese et al., 2006). In scenarios where priors are about spatial or temporal structure on hypotheses, *Independent Hypothesis Weighting* procedure was proposed by Ignatiadis et al. (2016), which clusters similar hypotheses into groups and assigns different weights to these groups. Hu et al. (2010) utilized the idea of both grouping and estimating the true null proportions within each group. Some more procedures in (Foygel-Barber & Candès, 2015; G'Sell et al., 2016; Lei & Fithian, 2016; Li & Barber, 2016a) incorporate a prior ordering as the auxiliary information to focus on more promising hypotheses near the top of the ordering. This motivation underlies also the first online multiple testing paper of (Foster & Stine, 2008).

Structure-adaptive BH algorithm (SABHA) (Li & Barber, 2016b) and *Adaptive p-value Thresholding* (AdaPT) (Lei & Fithian, 2018) are two recent FDR control adaptive methods which derive the feature vector dependent decision rules. SABHA first censors the p-values below a fixed level, and then uses the censored p-values to estimate the non-null proportion (using non-parametric methods in practice), and then applies the weighted BH procedure of (Genovese et al., 2006). AdaPT is based on adaptively estimating a Bayes-optimal p-value rejection threshold. At each iteration of AdaPT, an analyst proposes a significance threshold and observes partially censored p-values, then estimates the false discovery proportion (FDP) below the threshold, and proposes another threshold, until the estimated FDP is below the desired level.

The offline testing algorithm mostly related to our results is the *NeuralFDR* procedure proposed by Xia et al. (2017), which uses a neural network to parametrize the decision rule. This procedure in the offline setting, with access to all the p-values and the contextual features, comes up with a single decision rule $t(X)$ based on training a neural network for optimizing on the number of discoveries. In contrast, our method is in online multiple testing setup where we do not know all the p-values or the contextual features at once, and decision rules are different at each time, and varies as a function of previous outcomes and features.

1.2 Notation and Organization

We denote $[n] = \{1, \ldots, n\}$. Vectors are denoted by boldface letters. Given a sequence $(\Gamma_i)_{i \in \mathbb{N}}$, we denote by $\Gamma(n) = \sum_{i=1}^{n} \Gamma_i$ its partial sum. Bernoulli$(\cdot)$ and Uniform$(\cdot)$, with appropriate parameters, represent random variables drawn from the standard Bernoulli and uniform distributions respectively. $\mathcal{N}(\mu, \sigma^2)$ denotes a random variable from normal distribution with mean μ and variance σ^2.

Organization. The rest of the paper is organized as follows. In Section 2, we introduce the online multiple testing problem and survey some of the prior work in this area. In Section 3, we propose a new broad class

of powerful testing procedures which can incorporate contextual (side) information of the hypothesis, and present our theoretical guarantee of online FDR control for this class. In Section 4, we focus on a subclass of this above broad class, which are based on converting contextual features into weights, and using that to reweight the significance levels. In Section 5, we theoretically show the increase in power those can be obtained by using weighted procedures in the online setting. In Section 6, we design a practical algorithm for online multiple testing with contextual information, and demonstrate its performance on synthetic and real datasets.

2 Online False Discovery Rate Control: A Survey

We start with a review of the relevant notions in online multiple testing. The online model was first introduced by (Foster & Stine, 2008). Here we want to test an ordered (possibly infinite) sequence of hypotheses arriving in a stream, denoted by $\mathcal{H} = (H_1, H_2, H_3, \ldots,)$, where at each timestep t we have to decide whether to reject H_t having only access to previous information. As is standard notation, $H_t \in \{0, 1\}$ indicates if hypothesis t is *null* ($H_t = 0$) or *alternative* ($H_t = 1$). Each hypothesis is associated with a p-value P_t. The results in this paper do not depend on the actual test used for generating the p-value. By definition of a *valid* p-value, if the hypothesis H_t is *truly null*, then the corresponding p-value (P_t) is stochastically larger than the uniform distribution, i.e.,

$$\textbf{Super-uniformity of } P_t\textbf{:} \quad \text{If } H_t = 0 \ (H_t \text{ is null}) \text{ then } \Pr[P_t \le u] \le u, \quad \text{for all } u \in [0, 1]. \tag{1}$$

No assumptions are made on the marginal distribution of the p-values for hypotheses that are non-nulls (alternatives). Although they can be arbitrary, they should be stochastically smaller than the uniform distribution, since only then do they carry signal that differentiates them from nulls. Let $\mathcal{H}^0$ be the set $\{t \mid H_t = 0\}$ of indices of true null hypotheses in $\mathcal{H}$ and let $\mathcal{H}^1$ be the set $\{t \mid H_t = 1\}$ of remaining hypotheses in $\mathcal{H}$.

An online multiple testing procedure is defined as a *decision rule* which provides a sequence of significance level $\{\alpha_t\}$ and makes the following decisions:

$$R_t = \mathbb{1}\{P_t \le \alpha_t\} = \begin{cases} 1 & P_t \le \alpha_t & \Rightarrow \text{reject } H_t, \\ 0 & \text{otherwise} & \Rightarrow \text{accept } H_t. \end{cases} \tag{2}$$

A rejection of the null hypothesis H_t indicated by the event $R_t = 1$ is also referred as discovery.

Let us start by defining the false discovery rate (FDR), and true discovery rate (TDR) formally in the online setting. For any time T, denote the first T hypotheses in the stream by $\mathcal{H}(T) = (H_1, \ldots, H_T)$. Let $R(T) = \sum_{t=1}^{T} R_t$ be the total number of discoveries (rejections) made by the online testing procedure till time T, and let $V(T) = \sum_{t \in \mathcal{H}^0} R_t$ be the number of false discoveries (rejections) in them. Then online false discovery proportion (denoted as FDP) and the corresponding online false discovery rate (denoted as FDR) till time T are defined as:

$$\text{FDP}(T) := \frac{V(T)}{R(T) \vee 1}, \qquad \text{FDR}(T) := \mathbb{E}[\text{FDP}(T)].$$

Here, $R(T) \vee 1$ is the shorthand for $\max\{R(T), 1\}$. The expectation is over the underlying randomness. The false discovery rate is the expected proportion of false discoveries among the rejected hypotheses. As mentioned earlier, this criterion was first introduced by (Benjamini & Hochberg, 1995) in the offline multiple testing scenario. Similarly, let $S(T) = \sum_{t \in \mathcal{H}^1} R_t$ be the number of true discoveries (rejections) made by the online testing procedure till time T, and let $N_1(T)$ be the number of true alternatives (non-nulls). Then online true discovery proportion (denoted as TDP) and online true discovery rate (denoted as TDR) till time T are defined as:

$$\text{TDP}(T) := \frac{S(T)}{N_1(T) \vee 1}, \qquad \text{TDR}(T) := \mathbb{E}[\text{TDP}(T)].$$

The true discovery rate is also referred as to as *power*.

In the online setting, the question that arises is how can we generate a sequence of $(\alpha_1, \alpha_2, \ldots, \alpha_t, \ldots)$ such that we have control over the false discovery rate (FDR). Formally, the goal is to figure out a sequence of significance levels $(\alpha_t)_{t \in \mathbb{N}}$ such that the FDR can be controlled under a given level α at any time $T \in \mathbb{N}$, i.e.,

$$\sup_T \text{FDR}(T) \le \alpha.$$

Note that none of these above four metrics (FDP, FDR, TDP, or TDR) can be computed without the underlying true labels (ground truth).

A variant of FDR that arose from early works (Foster & Stine, 2008) on online multiple hypothesis testing is that of *marginal FDR* (denoted as mFDR), defined as:

$$\text{mFDR(T)}_\eta = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)] + \eta}.$$

A special case of mFDR(T) $= \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)]+1}$ is when $\eta = 1$. mFDR measures the ratio of expected number of false discoveries to expected number of discoveries. Informally, while FDR controls a property of the realized set of tests, mFDR is the ratio of the two expectations over many realizations. In general, the gap between FDR and mFDR can be very significant, and in particular controlling mFDR does not ensure controlling FDR at a similar level (Javanmard & Montanari, 2018). We will also provide a theoretical guarantee on mFDR control in a contextual setting under some weaker assumptions on p-values.

Generalized Alpha-Investing Rules. Foster & Stine (2008) proposed the first class of online multiple testing rules (referred to as alpha-investing rules) to control mFDR (under some technical assumptions). Aharoni & Rosset (2014) further extended this class to generalized alpha-investing (GAI) rules, again to control mFDR. The general idea behind these rules is that you start with an initial wealth, which is under the desired control level. Some fraction of the wealth is spent (i.e., wealth decreases) for testing each hypothesis. However, each time a discovery occurs, a reward is earned, i.e., wealth increases for further tests. Building on these results, Javanmard & Montanari (2015) demonstrated that monotone GAI rules, appropriately parameterized, can control the (unmodified) FDR for independent p-values. We now introduce the GAI rules conceptualized in these works.

Given a sequence of input p-values $(P_1, P_2, \ldots)$, a generalized alpha-investing rule generates a sequence of significance levels $(\alpha_1, \alpha_2, \ldots)$, which is then used for generating the sequence of decisions R_t's as in (2). In a GAI rule, the significance levels at time t is a function of prior decisions (till time $t-1$):

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}).$$

In particular, this means that α_t does not directly depend on the observed p-values but only on past decisions. Let $\mathcal{F}^t = \sigma(R_1, \ldots, R_t)$ denote the sigma-field of decisions till time t. In GAI rules, we insist that $\alpha_t \in \mathcal{F}^{t-1}$.

Any GAI rule begins with an wealth of $W(0) > 0$, and keeps track of the wealth $W(t)$ available after t testing steps. Formally, a generalized alpha-investing rule is specified by three (sequences of) functions $\alpha_t, \phi_t, \psi_t \in \mathcal{F}^{t-1}$. Here, ϕ_t is the penalty of testing a new hypothesis H_t, and ψ_t is the reward for making a rejection (discovery) at time t. In other words, a part of the wealth is used to test the tth hypothesis at level α_t, and the wealth is immediately decreased by an amount ϕ_t. If the tth hypothesis is rejected, that is if $R_t = 1$, then an extra wealth equaling an amount ψ_t is added to the current wealth. This can be explicitly stated as:

$$W(0) = w_0, \tag{3}$$

$$W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t, \tag{4}$$

where $0 < w_0 < \alpha$ is the initial wealth. The parameters w_0 and the sequences α_t, ϕ_t, ψ_t are all user-defined. A requirement is that the total wealth $W(t)$ is always non-negative, and hence $\phi_t \leq W(t-1)$. If the wealth ever equals zero, the procedure is not allowed to reject any more hypotheses since it has to set $\alpha_t = 0$ from then on. An additional restriction is needed from the goal to control FDR, in that whenever a rejection takes place, ψ_t cannot be arbitrary. Formally, this constraint is defined as:

$$\psi_t \leq \min\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\}. \tag{5}$$

Javanmard & Montanari (2015) defined b_t as a user-defined constant B_0, setting it to $\alpha - w_0$. Recently, Ramdas et al. (2017b) demonstrated that setting

$$b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t - 1\}$$

could potentially lead to larger statistical power. Here, for a positive integer k,

$$\rho_k := \min_{i \in \mathbb{N}} \{\sum_{t=1}^{i} R_t = k\},$$

is the time of kth rejection (discovery). Ramdas et al. (2017b) refer to this particular setting of b_t as GAI++ rules. Unless otherwise specified, we use this improved setting of b_t throughout this paper.

Another important property when dealing with GAI rules is that of monotonicity. As studied by (Javanmard & Montanari, 2015, 2018) in the context of GAI rules, and as is predominantly the case in offline multiple testing, monotonicity is the following condition:

$$\text{if } \tilde{R}_i \leq R_i \text{ for all } i \leq t - 1 \text{ then } \quad \alpha_t(\tilde{R}_1, \ldots, \tilde{R}_{t-1}) \leq \alpha_t(R_1, \ldots, R_{t-1}). \tag{6}$$

While these online FDR procedures are widely used, a major shortcoming of them is that they ignore additional information that is often available during testing. Each hypothesis, in addition to the p-value, could have a feature vector which encodes contextual information related to the tested hypothesis. For example, in genetic association studies, each hypothesis tests the correlation between a variant and the trait. We have a rich set of features for each variant (e.g., its location, conservation, epigenetics, etc.) which could inform how likely the variant is to have a true association.

In this paper, we introduce the problem contextual online multiple testing, which captures the presence of this auxiliary information in modern applications.

Level based On Recent Discovery (LORD) Rules. One specific subclass of GAI rules (proposed by (Javanmard & Montanari, 2015, 2018)) that arguably is state-of-the-art in online multiple hypothesis testing and performs consistently well in practice is known as *Level based On Recent Discovery* (LORD). We will focus on weighted variants of LORD later in this paper when we discuss about statistical power of tests. In fact, we will consider the recently improved LORD++ rules (proposed by (Ramdas et al., 2017b)) that achieves the same or better statistical power than the LORD rules (uniformly).

The idea behind LORD (and LORD++) rules is that the significance level α_t is a function based only on *most recent discovery time*. Formally, we start with any sequence of nonnegative numbers $\gamma = (\gamma_t)_{t=1}^{\infty}$, which is monotonically non-increasing with $\sum_{t=1}^{\infty} \gamma_t = 1$. At each time t, let τ_t be the last time a discovery was made before t, i.e.,

$$\tau_t := \max\{i \in \{1, \ldots, t-1\} : R_i = 1\},$$

with $\tau_t = 0$ for all t before the first discovery. The LORD rule defines α_t, ϕ_t, ψ_t in the following generalized alpha-investing fashion.[2]

---

[2]Note that Javanmard & Montanari (2018) defined three versions of LORD that slightly vary in how they set the significance levels. In this paper, we stick to one version, though much of the discussion in this paper also holds for the other versions.

Level based On Recent Discovery (LORD) (Javanmard & Montanari, 2015, 2018):

$$W(0) = w_0,$$

$$\phi_t = \alpha_t = \begin{cases} \gamma_t w_0 & \text{if } t \leq \rho_1 \\ \gamma_{t-\tau_t} b_0 & \text{if } t > \rho_1, \end{cases}$$

$$\psi_t = b_0,$$

$$b_0 = \alpha - w_0.$$

Typically, we will set $w_0 = \alpha/2$, in which case, the above rule could be simplified as $\phi_t = \alpha_t = \gamma_{t-\tau_t} b_0 = \gamma_{t-\tau_t} \alpha/2$.

As with any GAI rule, (Ramdas et al., 2017b) showed that one could replace b_0 with $b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t - 1\}$ to achieve potentially better power, while still achieving online FDR control at level α. With this replacement, we defined LORD++ as follows.

Improved Level based On Recent Discovery (LORD++) (Ramdas et al., 2017b):

$$W(0) = w_0 \geq \alpha/2,$$

$$\phi_t = \alpha_t = \gamma_{t-\tau_t} b_t,$$

$$\psi_t = b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t - 1\}.$$

It can be easily observed that both LORD and LORD++ rules satisfy the monotonicity condition from (6).

SAFFRON Procedure. This is a very recently proposed online FDR control procedure by Ramdas et al. (2018). The main difference between SAFFRON (Serial estimate of the Alpha Fraction that is Futilely Rationed On true Null hypotheses) and the previously discussed LORD/LORD++ procedures comes in that SAFFRON is an adaptive method, based on adaptively estimating the proportion of true nulls. SAFFRON can be viewed as an online extension of Storey's adaptive version of BH procedure in the offline setting. SAFFRON does not belong to the GAI class, whose extension to the contextual online setting is the main focus of this paper. See Appendix C for more details about SAFFRON, and the experiments with SAFFRON that suggests that contextual information could potentially help here too.

3   Contextual Online FDR Control

We start with an informal definition of the contextual online multiple testing problem. Consider a setting where we test an ordered (possibly infinite) sequence of null hypotheses, denoted $\mathcal{H} = (H_1, H_2, H_3, \ldots,)$. Each hypothesis H_t is associated with a p-value $P_t \in (0, 1)$ and a vector of contextual features $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$, thus can be represented by a tuple (H_t, P_t, X_t). We observe P_t and X_t, but do not know H_t. The goal of contextual online testing, is at each step t, decide whether to reject H_t having only access to previous decisions and contextual information seen so far. The overall goal is to control online FDR under a given level α at any time and improve the number of useful discoveries by using the contextual information. All missing details in this section are collected in Appendix A.

Under the alternative, we denote the density distribution (PDF) of p-values as $f_1(p \mid X)$ (for $X \in \mathcal{X}$), and the cumulative distribution (CDF) of p-values as $F_1(p \mid X)$. Here $f_1(p \mid X)$ can be any *arbitrary unknown* function, as long as the p-values are stochastically smaller than those under the null. Note that $f_1(p \mid X)$ is not identifiable from the data as we never observe H_t's directly. This can be illustrated through a simple example described in Appendix A. Let us now formally define the contextual online FDR control problem.

Definition 1 (Contextual Online FDR Control Problem). *Given a (possibly infinite) sequence of (P_t, X_t)'s $(t \in \mathbb{N})$ where $P_t \in (0,1)$ and $X_t \in \mathcal{X}$, the goal is to generate a significance levels $\alpha_t's$ as a function of prior decisions and contextual features*

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t),$$

and a corresponding set of decisions

$$R_t = \begin{cases} 1 & P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t), \\ 0 & otherwise. \end{cases}$$

such that $\sup_T FDR(T) \le \alpha$.

A related definition would be to control mFDR (instead of FDR) under level α. Note that in the contextual setting, we consider the significance levels to be functions of prior decisions and contextual features seen so far. This differs from existing online multiple testing rules described in the previous section where the significance levels (α_t's) depend only on prior results ($\alpha_t \in \mathcal{F}^{t-1}$), i.e., does not use any contextual information.

In the presence of contextual information, we use the sigma-field of decisions till time t as $\mathcal{F}^t = \sigma(R_1, \ldots, R_t)$, and the sigma-field of features till time t as $\mathcal{G}^t = \sigma(X_1, \ldots, X_t)$. Our first contribution in this paper is to define a contextual extension of GAI rules, that we refer to as *Contextual Generalized Alpha-Investing* (contextual GAI or CGAI) rules. A contextual GAI rule is defined through three functions, $\alpha_t, \phi_t, \psi_t \in \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$, that are all computable at time t.

A valid contextual GAI rule is required to satisfy the following conditions:

Contextual GAI:

$$\textbf{Initial wealth: } W(0) = w_0, \text{with } 0 < w_0 < \alpha, \tag{7}$$
$$\textbf{Wealth Update: } W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t, \tag{8}$$
$$\textbf{Non-negativity: } \phi_t \le W(t-1), \tag{9}$$
$$\textbf{Upper bound on reward: } \psi_t \le \min\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\}. \tag{10}$$

We set $b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t-1\}$ as proposed by (Ramdas et al., 2017b).

Also as with GAI rules, we can add a monotonicity property to a contextual GAI rule as follows:

Monotoncity: If $\tilde{R}_i \le R_i$ for all $i \le t-1$ then

$$\alpha_t(\tilde{R}_1, \ldots, \tilde{R}_{t-1}, X_1, \ldots, X_t) \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t), \text{ for any fixed } \mathbf{X}^t = (X_1, \ldots, X_t). \tag{11}$$

A contextual GAI rule satisfying the monotonicity condition is referred to as *monotone contextual GAI*.

Our first result establishes the online FDR control for any monotone contextual GAI rule under independence of p-values, and the mutual independence of p-values and contextual features under the null.[3]

We start by presenting the following lemma, which is an intermediate result for the proof of FDR later. Note that $R_t = \mathbb{1}\{P_t \le \alpha_t\}$, where $\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)$ is $\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$-measurable and is a coordinatewise non-decreasing function of $R_1, \ldots, R_{t-1}$ for any fixed $\mathbf{X}^t = (X_1, \ldots, X_t)$. Due to the super-uniformity of p-values under the null (1), we immediately have that for independent p-values under the null the following conditional super-uniformity condition holds.

$$\Pr[P_t \le \alpha_t \mid \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)] \le \alpha_t. \tag{12}$$

---

[3]Note that a standard assumption in hypothesis testing is that the p-values under the null are uniformly distributed in $(0, 1)$, which implies the mutual independence of p-values and contextual features under the null.

Lemma 1 states a more general result about super-uniformity of independent p-values under the null. These independence assumptions are standard in multiple testing literature (see, e.g., Javanmard & Montanari (2018); Ramdas et al. (2017b); Xia et al. (2017) among others). The proof is based on a *leave-one-out* technique a variant of which was also used by Ramdas et al. (2017b) (and also by Javanmard & Montanari (2018)) in their analyses. The main distinction for us comes in that we consider the sigma-field at each time t as $\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$ including the information of contextual features till time t, instead of just $\mathcal{F}^{t-1}$.

Lemma 1 (Super-uniformity). *Let $g : \{0,1\}^T \to \mathbb{R}$ be any coordinatewise non-decreasing function such that $g(\mathbf{R}) > 0$ for any vector $\mathbf{R} \neq (0, \dots, 0)$. Then for any index $t \leq T$ such that $t \in \mathcal{H}^0$, we have*

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \dots, R_{t-1}, X_1, \dots, X_t)\}}{g(R_1, \dots, R_T) \vee 1}\middle|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \leq \mathbb{E}\left[\frac{\alpha_t(R_1, \dots, R_{t-1}, X_1, \dots, X_T)}{g(R_1, \dots, R_T) \vee 1}\middle|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right].$$

The following theorem establishes the FDR property on any monotone contextual GAI rule under some independence assumptions on the p-values. Note that, as mentioned above, the p-values (P_t) could be arbitrary related to the contextual features (X_t) under the alternative (when $H_t = 1$). The proof considers two cases: (a) $H_t = 0$ in which case it uses the super-uniformity statement established in Lemma 1 and the fact that $\psi_t \leq \phi_t/\alpha_t + b_t - 1$ (by definition), and (b) $H_t = 1$ in which case we use the fact that $\psi_t \leq \phi_t + b_t$ (by definition).

Theorem 1. *[FDR Control] Consider a sequence of $((P_t, X_t))_{t\in\mathbb{N}}$ of p-values and contextual features. If the p-values P_t's are independent, and additionally P_t are independent of all $(X_t)_{t\in\mathbb{N}}$ are independent under the null (whenever $H_t = 0$), then for any monotone contextual generalized alpha-investing rule (i.e., satisfying conditions (7), (8), (9), (10), and (11)), we have online FDR control,*

$$\sup_{T\in\mathbb{N}} \mathrm{FDR}(T) \leq \alpha.$$

Turning our attention to mFDR, we can prove a guarantee for mFDR control under a weaker condition than that in Theorem 1. In particular, one can relax the independence assumptions to a weaker conditional super-uniformity assumption from (12) (as mentioned above, the validity of p-values (1) and independence implies conditional super-uniformity). We define that a null p-value is conditionally super-uniform on past discoveries and contextual features so far, if the following holds

Conditional super-uniformity of P_t: If $H_t = 0$ (H_t is null) then $\Pr[P_t \leq \alpha_t \mid \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)] \leq \alpha_t$. (13)

Our next theorem formalizes the mFDR control for any contextual GAI rule (not necessarily monotone) under the above condition.

Theorem 2. *[mFDR Control] Consider a sequence of $((P_t, X_t))_{t\in\mathbb{N}}$ of p-values and contextual features. If the p-values P_t's are conditionally super-uniform distributed (as in (13)), then for any contextual generalized alpha-investing rule (i.e., satisfying conditions (7), (8), (9), and (10)), we have online mFDR control,*

$$\sup_{T\in\mathbb{N}} \mathrm{mFDR}(T) \leq \alpha.$$

Remark 1. *For arbitrary dependent p-values and contextual features, the FDR control can also be obtained by using a modified LORD rule defined in Javanmard & Montanari (2018), under a special case where the contextual features are transformed into weights satisfying certain conditions. See Proposition 2 for a formal statement.*

4 Context-weighted Generalized Alpha-Investing Rules

The contextual GAI rules, introduced in the previous section, form a very general class of online multiple testing rules. In this section, we focus on a subclass of these rules, which we refer to as *Context-weighted Generalized Alpha-Investing* (context-weighted GAI or CwGAI) rules. Specifically, we consider α_t to be a product of two functions with the first one of previous decisions and second one based on the current contextual feature. Borrowing from offline multiple testing literature, we refer to the second parametric function as a *weight function* (denoted by $\omega(\cdot)$). Therefore, a context-weighted GAI rule has this form,

$$\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t) := \alpha_t(R_1, \ldots, R_{t-1}) \cdot \omega(X_t; \theta), \tag{14}$$

where $\omega(X_t; \theta)$ is a parametric function with a parameter set $\theta \in \Theta$. Since CwGAI is a subclass of CGAI rules, the above FDR and mFDR control theorems are valid for this class too.

Our reasons for considering this subclass of contextual GAI rules include: *a)* With context-weighted generalized alpha-investing rules we obtain a simpler form of α_t by separating the contextual features from that of previous outcomes, making it easier to design functions that satisfy the monotonicity requirement of the GAI rules, *b)* It is now convenient to model the weight function by any parametric function, and *c)* We can learn the parameters of the weight function empirically by maximizing the number of discoveries. This forms the basis of a practical algorithm for contextual online FDR control that we describe in Section 6. Note that the GAI rules are still a subclass of context-weighted generalized alpha-investing rules by taking the weight function as a constant 1. We illustrate the relationship among various classes of testing rules in Figure 1.

The idea of weighting p-values by using prior information has been widely studied in offline multiple testing setup (Genovese et al., 2006; Ignatiadis et al., 2016; Lei & Fithian, 2018; Li & Barber, 2016b; Ramdas et al., 2017a; Xia et al., 2017). In many applications, contextual information can provide some prior knowledge about the true underlying state at current time, which may be incorporated in by a weight $\omega_t = \omega(X_t; \theta)$. Intuitively, the weights indicate the strength of a prior belief whether the underlying hypothesis is null or not. Therefore, a larger weight $\omega_t > 1$ provides more belief of a hypothesis being an alternative which makes the procedure to reject it more aggressively, while a smaller weight $\omega_t < 1$ indicates a higher likelihood of a null which makes the procedure reject it more conservatively.

Weighting in Online vs. Offline Setting with FDR Control. In the offline setting with a fixed set $\{P_1, \ldots, P_n\}$ of p-values, the idea of using prior weights is as follows: a) first the given set of weights (say, $\hat{\omega}'_i$s) are rescaled to make the average equal to one giving rise to new ω_i's ($i \in [n]$), and b) then any existing offline FDR control algorithm is applied by replacing P_i with the weighted p-value P_i/ω_i. This p-value weighting in the offline setting was first suggested by (Genovese et al., 2006), and we refer the reader to that paper for a detailed account on p-value weighting. However, in an online setting, the weights ω_t's are computed at each timestep t without access to the total number of hypothesis or contextual information ahead of time, which means weights cannot be rescaled to have unit mean in advance.

Instead of weighting the p-values, as represented in (14), we consider weighting the significance levels α_t's, as was also considered by Ramdas et al. (2017a). Note that weighting p-values is equivalent to weighting significance levels in terms of decision rules conditioning on the same significance levels, i.e., given the same α_t's, we have $\{P_t/\omega_t \le \alpha_t\} \equiv \{P_t \le \alpha_t \omega_t\}$ for all t. The main difference is that when α_t's are weighted, the penalty ϕ_t's and rewards ψ_t's also needs to be adjusted according to the GAI constraints. For example, as dictated by (5), if we overstate our prior belief in the hypothesis being alternative by assigning a large $\omega_t > 1$, the penalty will need to be more or the reward will need to be less.

Context-weighted LORD++. Next let us see how the principle of context weighting can be applied to LORD++ rules defined in Section 2. Given a weight function, $\omega : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$, we define the context-

GAI: Generalized Alpha Investing

CGAI: Contextual GAI

CwGAI: Context-weighted GAI

LORD: Level based On Recent Discovery
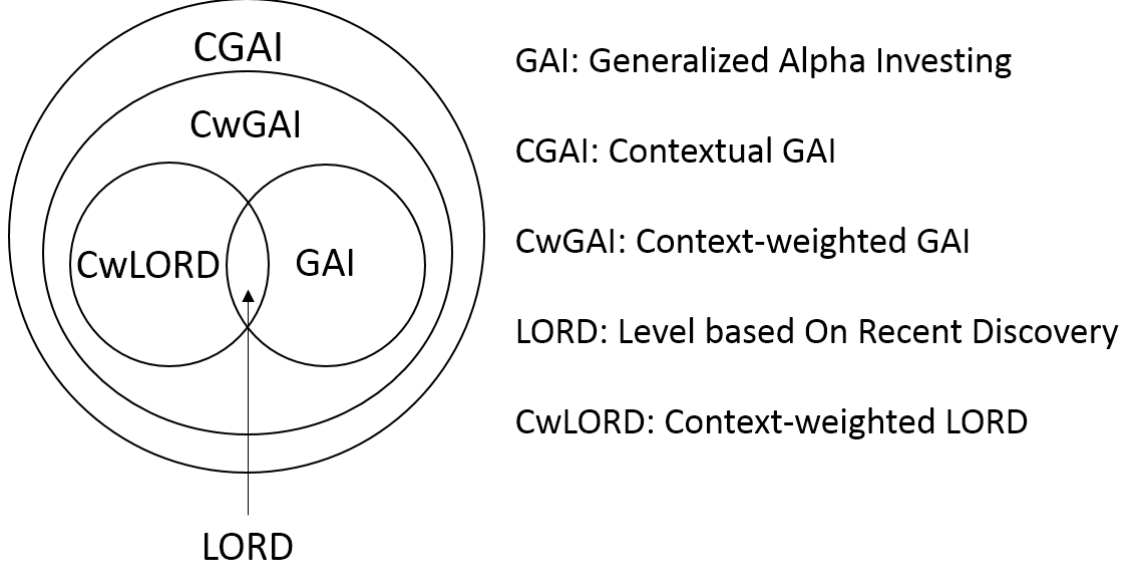
CwLORD: Context-weighted LORD

Figure 1: Relationship among various testing rules. One could replace LORD with LORD++ (resp. CwLORD with CwLORD++).

weighted LORD++ (CwLORD++) testing rule as follows.

> Context-weighted LORD++ (CwLORD++):
>
> $$W(0) = w_0,$$
> $$\phi_t = \alpha_t = \min\{\gamma_{t-\tau_t} b_t \cdot \omega(X_t; \theta), W(t-1)\},$$
> $$\psi_t = b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t - 1\}.$$

Similarly, given the function ω, we can also define context-weighted LORD using the LORD rules from Section 2. As pointed out before, when α_t's are reweighted, the penalty ϕ_t's are also adjusted accordingly. This provides a clean way of incorporating the weights in an online setup without having to rescale the weights to have unit mean.

   In the next section, we see how weighting could be beneficial in an online setting. In Section 6, we discuss an online multiple testing procedure based on CwLORD++ (see Algorithm 1), where we use a neural network to model the weight function ω with parameters trained to adaptively increase the discovery proportion over time. Our experimental investigation will be based on this procedure.

5   Statistical Power of Weighted Online Rules

In this section, we answer the question whether weighting helps in an online setting in terms of increased power. We answer this question in affirmative, in the context of the popular LORD procedure of Javanmard & Montanari (2018). The benefits of weighting in the offline setting, in terms of increased power was first studied by Genovese et al. (2006), who showed that a weighted BH procedure improves the power over the corresponding unweighted procedure if weighting is *informative*, which roughly means that the weights are positively associated with the non-nulls. Missing details from section are collected in Appendix B.

   We use a model setup as in (Genovese et al., 2006) modified to an online setting. In particular, we consider a mixture model where each null hypothesis is false with probability π_1 independently of other

hypotheses, and the p-values corresponding to different hypotheses are all independent. While the mixture model is admittedly *idealized*, it does offer a natural ground for comparing the power of various testing procedures (Genovese et al., 2006; Javanmard & Montanari, 2018). Let us start with a formal description of the mixture model. The rest of the discussion in this section will be with respect to this mixture model.

Mixture Model. For any $t \in \mathbb{N}$, let

$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_1, \ldots, X_t \overset{\text{i.i.d.}}{\sim} \mathcal{L}(\mathcal{X}),$$
$$P_t \mid H_t = 0, X_t \sim \text{Uniform}(0, 1),$$
$$P_t \mid H_t = 1, X_t \sim F_1(p \mid X_t).$$

where $0 < \pi_1 < 1$ and where $\mathcal{L}(\mathcal{X})$ is a probability distribution on the contextual feature space $\mathcal{X}$. Let $F = \int F_1(p \mid X) \mathrm{d}\mathcal{L}(\mathcal{X})$ be the marginal distribution of p-value under alternative, which we assume is stochastically smaller than the uniform distribution. Marginally, the p-values are i.i.d. from the CDF $G(a) = (1 - \pi_1)U(a) + \pi_1 F(a)$, where $U(a)$ is the CDF of Uniform(0,1). For this mixture model, a lower bound (and in some cases tight) bounds on the power of the LORD procedure was established by Javanmard & Montanari (2018). Our aim will be to compare this with the power of a weighted version of LORD to establish a separation.

We consider the model of general weighting where weight is a random variable that are conditionally independent of P_t given H_t for all $t = 1 \ldots, \infty$. We assume that weight ω has different marginal distributions under null and alternative,

$$\omega \mid H_t = 0 \sim Q_0, \quad \omega \mid H_t = 1 \sim Q_1, \tag{15}$$

for some unknown probability distributions Q_0, Q_1 that are continuous on $(0, \infty)$.

Remark 1. *Note that this framework of weighting in (15) is very general. For example, it includes as a special case, the following contextual weighting scenario, where we assume that there exists two different implicit weight functions of contextual features X, $\omega_0(X)$ and $\omega_1(X)$, and the resulting marginal distributions of weights are computed as:*

$$\omega \mid H_t = 0 \sim \int \omega_0(X) \mathrm{d}\mathcal{L}(\mathcal{X}), \quad \omega \mid H_t = 1 \sim \int \omega_1(X) \mathrm{d}\mathcal{L}(\mathcal{X}).$$

So, here we can think of the distributions Q_0 and Q_1 in (15) as $\int \omega_0(X) \mathrm{d}\mathcal{L}(\mathcal{X})$ and $\int \omega_1(X) \mathrm{d}\mathcal{L}(\mathcal{X})$, respectively.

Under (15), the marginal distribution of ω is $Q = (1 - \pi_1)Q_0 + \pi_1 Q_1$. For $j = 0, 1$, let $u_j = \mathbb{E}[\omega \mid H_t = j]$ be the means of Q_0 and Q_1 respectively. We also assume that the following notion of informative-weighting, based on a similar condition used by (Genovese et al., 2006) in the offline setting.

> **Informative-weighting:** $u_0 < 1, \quad u_1 > 1, \quad u = \mathbb{E}[\omega] = (1 - \pi_1)u_0 + \pi_1 u_1 = 1.$ (16)

Remark 2. *We would like to point that $\mathbb{E}[\omega] = 1$ is an assumption on the weight distribution that may not hold in a practical setting. So for empirical experiments, we will use an instantiation of CwLORD++ (see Section 6.1), that does not require this assumption. But for the theoretical comparison of the power of different procedures, we make this assumption, under which we can use the p-value reweighting akin to the offline setting.*

Let the weights ω_t's be random variables that are drawn i.i.d. from this mixture model with marginal distribution Q. Taking the LORD++ rule from Section 2, we define a weighted LORD++ rule as follows.

Definition 2 (Weighted LORD++). *Given a sequence of p-values, $(P_1, P_2, \dots)$ and weights $(\omega_1, \omega_2, \dots)$, apply LORD++ with level α to the weighted p-values $(P_1/\omega_1, P_2/\omega_2, \dots)$.*

We want to emphasize that the weighted LORD++ rule actually reweights the p-values (as done in the offline weighted BH procedure (Genovese et al., 2006)) and then applies the original LORD++ to these reweighted p-values. So this is slightly different from the idea of CwLORD++, that we mentioned above, which reweights the significance levels and then applies it to the original p-values. To understand the difference, let us start from their definitions.

In LORD++, the penalty is $\phi_t = \alpha_t = \gamma_{t-\tau_t} b_t < W(t-1)$, which is always less than current wealth due to the construction of γ_t's and w_0. So in weighted LORD++, we are comparing reweighted p-value P_t' to the level $\alpha_t = \gamma_{t-\tau_t} b_t$, which is equivalent to comparing the original p-value P_t to the level $\alpha_t' = \gamma_{t-\tau_t} b_t \omega_t$.

On the other hand, in CwLORD++, we take a penalty of the form $\tilde{\phi}_t = \tilde{\alpha}_t = \min\{\gamma_{t-\tau_t} b_t \omega_t, W(t-1)\}$. We take the minimum of reweighted significance level and the current wealth, to prevent the penalty $\gamma_{t-\tau_t} b_t \omega_t$ from exceeding the current wealth which would violate a tenet of the alpha-investing rules.

A simple corollary is that the actual significance levels used in weighted LORD++ are greater than those in CwLORD++, i.e.,

$$\alpha_t' = \gamma_{t-\tau_t} b_t \omega_t \geq \min\{\gamma_{t-\tau_t} b_t \omega_t, W(t-1)\} = \tilde{\alpha}_t.$$

That implies the power of weighted LORD++ is equal to or greater than the power of CwLORD++, whereas the FDR of weighted LORD++ may also be higher than that of CwLORD++. From Theorem 1, we know that we have FDR control with CwLORD++, however that result does not hold for weighted LORD++ (as weighted LORD++ is not strictly a contextual GAI rule) We now show that the above weighted LORD++ can still control online FDR at any given level α under the condition $\mathbb{E}[\omega] = 1$, which we do in the following proposition.

Proposition 1. *Suppose that the weight distribution satisfies the informative-weighting assumption in (16). Suppose that p-values P_t's are independent, and are conditionally independent of the weights ω_t's given H_t's. Then the weighted LORD++ rule can control the online FDR at any given level α, i.e.,*

$$\sup_{T \in \mathbb{N}} \text{FDR}(T) \leq \alpha.$$

Weakening the Assumptions from Proposition 1. In most applications, the independence between p-values and weights needed in Proposition 1 is not guaranteed. Javanmard & Montanari (2018) achieved the FDR control under dependent p-values by using a modified LORD rule, which sets $\psi_t = b_0$ and $\alpha_t = \phi_t = \gamma_t W(\tau_t)$ with the fixed sequence (γ_t) satisfying $\sum_{t=1}^{\infty} \gamma_t (1 + \log(t)) \leq \alpha/b_0$.

We can extend the FDR control results to the dependent weighed p-values. In particular, as long as the following condition is satisfied, i.e., for each weighted p-value (P_t/ω_t) marginally

$$\Pr[P_t/\omega_t \leq u \mid H_t = 0] \leq u, \quad \text{for all } u \in [0, 1], \tag{17}$$

then the upper bound of FDR stated in Theorem 3.7 in Javanmard & Montanari (2018) is valid for weighted p-values. Specifically, if the modified LORD rule in Example 3.8 of Javanmard & Montanari (2018) is applied to the weighted p-values under the assumption in (17), then the FDR can be controlled below level α.

We formally state the results in the following proposition.

Proposition 2. *Suppose that the weight distribution satisfies the informative-weighting assumption in* (16). *And weighted p-values* (P_t/ω_t) *marginally satisfy* (17). *Then the modified LORD++ rule that applies to the weighted p-values can control the online FDR at any given level α, i.e.,*

$$\sup_{T\in\mathbb{N}} \text{FDR}(T) \le \alpha.$$

The proofs of these extension are almost the same as those in Javanmard & Montanari (2018) and are omitted here.

Lower Bound on Statistical Power of Weighted LORD++. In order to compare different procedures, it is important to estimate their statistical power. Here, we analyze the power of the weighted LORD++. Define $D(a) = \Pr[P/\omega \le a]$ as the marginal distribution of weighted p-values. Under the assumptions on the weight distribution from (15), the marginal distribution of weighted p-value equals,

$$
\begin{aligned}
D(a) = \Pr[P/\omega \le a] &= \int \Pr[P/\omega \le a \mid \omega = w]\, \mathrm{d}Q(w) \\
&= \int \sum_{h\in\{0,1\}} \Pr[P/\omega \le a \mid \omega = w, H = h]g(h \mid w)\, \mathrm{d}Q(w) \\
&= \int \sum_{h\in\{0,1\}} \Pr[P/w \le a \mid H = h]g(h \mid w)\, \mathrm{d}Q(w) \\
&= \int \sum_{h\in\{0,1\}} ((1-h)aw + hF(aw))g(h \mid w)\, \mathrm{d}Q(w) \\
&= \int \sum_{h\in\{0,1\}} ((1-h)aw + hF(aw))\, \mathrm{d}Q(w \mid h)g(h) \\
&= \sum_{h\in\{0,1\}} \int ((1-h)aw + hF(aw))\, \mathrm{d}Q(w \mid h)g(h) \\
&= (1-\pi_1) \int aw\, \mathrm{d}Q(w \mid h = 0) + \pi_1 \int F(aw)\, \mathrm{d}Q(w \mid h = 1) \\
&= (1-\pi_1)\mu_0 a + \pi_1 \int F(aw)\, \mathrm{d}Q_1(w).
\end{aligned}
\tag{18}
$$

Theorem 3. *Let $D(a) = \Pr[P/\omega \le a]$ be the marginal distribution of weighted p-values as in* (18). *Then, the average power of weighted LORD++ rule is almost surely bounded as follows:*

$$\liminf_{T\to\infty} \text{TDR}(T) \ge \Big(\sum_{m=1}^{\infty} \prod_{j=1}^{m}(1 - D(b_0\gamma_j)) \Big)^{-1}.$$

The proof of Theorem 3 uses the similar technique as the proof of the statistical power of LORD in (Javanmard & Montanari, 2018). The main distinction is that we replace the marginal distribution of p-values by the marginal distribution of weighted p-values. Theorem 3 also holds for a weighted LORD procedure (replacing LORD++ by LORD in Definition 2).

Comparison of Power. Next, we establish conditions under which a weighting could lead to increased power for LORD. We work with (a version of) the popular LORD procedure from Javanmard & Montanari (2018), which sets

$$W(0) = w_0 = b_0 = \alpha/2, \;\; \phi_t = \alpha_t = b_0\gamma_{t-\tau_t}, \;\; \psi_t = b_0. \tag{19}$$

As shown by Javanmard & Montanari (2018), the average power of LORD, under the mixture model, almost surely equals[4]

$$\text{For LORD: } \liminf_{T \to \infty} \text{TDR}(T) = \Big(\sum_{m=1}^{\infty} \prod_{j=1}^{m} (1 - G(b_0 \gamma_j)) \Big)^{-1}, \tag{20}$$

where $G(a) = (1 - \pi_1) U(a) + \pi_1 F(a)$ as defined earlier.

Definition 3 (Weighted LORD). *Given a sequence of p-values, $(P_1, P_2, \dots)$ and weights $(\omega_1, \omega_2, \dots)$, apply LORD* (19) *with level α to the weighted p-values $(P_1/\omega_1, P_2/\omega_2, \dots)$.*

The FDR control of weighted LORD under an independence assumption on p-values and an assumption (16) on weights follows from Proposition 1. From the proof of Theorem 3, the average power of weighted LORD almost surely equals

$$\text{For weighted LORD: } \liminf_{T \to \infty} \text{TDR}(T) = \sum_{m=1}^{\infty} \prod_{j=1}^{m} (1 - D(b_0 \gamma_j))^{-1}. \tag{21}$$

Assume F is differentiable and let $f = F'$ be the PDF of p-values under alternative. Due to the fact that p-values under alternative are stochastically dominated by the uniform distribution, there exists some $a_0 > 0$ such that $f(a) > 1$ for all $0 \leq a < a_0$. The following theorem is based on comparing this power on weighted LORD from (21) with the power on LORD from from (20).

Theorem 4. *[Power Separation] Suppose that the parameters in LORD* (19) *satisfy $b_0 \gamma_1 < a_0$, and the weight distribution satisfies $\Pr[\omega < a_0/(b_0 \gamma_1) \mid H_t = 1] = 1$ for every $t \in \mathbb{N}$ and the informative-weighting assumption in* (16). *Then, the average power of weighted LORD is greater than equal to that of LORD almost surely.*

Note that the general weighting here includes the context-weighting discussed in Section 4 (through Remark 1). So the results here indicate that using the informative context-weighting in the LORD rules will help in making more true discoveries.

Remark 3. *Intuitively, the condition implies that to achieve higher power while controlling FDR, the weights given to alternate hypotheses cannot be too large. Let us discuss this point in the context of weighted LORD++ and context-weighted LORD++.*

In weighted LORD++, we can always assign large weights to make the reweighted p-values small enough to be rejected, in order to achieve high power. But this can lead to a loss in FDR control which is why we need the restriction of $\mathbb{E}[\omega] = 1$ for proving the FDR control in Proposition 1. Therefore, it is natural to have weights not too large.

If we consider reweighting the significance levels as in CwLORD++, assigning a large weight will not affect the FDR control (we prove that FDR is controlled for any choice of weights in Theorem 1). However, the price is paid in terms of power. When we use a large weight, the penalty ϕ_t increases and therefore the wealth might go quickly down to zero. Once the wealth is exhausted, the significance levels afterwards must all be zero and thus preventing any further discoveries.

We now conclude this discussion with a simple example of how the conditions of Theorem 4 are easily satisfied in a common statistical model.

---

[4]Javanmard & Montanari (2018) proposed multiple versions of LORD, and as noted by them the bound in (20) lower bounds the average power on all the versions of LORD for the above mixture model.

Example 1. *To further interpret the weight condition in Theorem 4, let us take a concrete example. Consider the hypotheses $(H_1, \ldots, H_T)$ concerning the means of normal distributions (referred to as* normal means *model). This model corresponds to getting test statistics $Z_t \sim \mathcal{N}(\mu, 1)$. So the two-sided p-values are $P_t = 2\Phi(-|Z_t|)$, where Φ is the CDF of standard normal distribution. Suppose under the null hypothesis $\mu = 0$, and under the alternative hypothesis $0 < \mu \leq 4$. Then from simple computation we obtain that $a_0 > 0.022$ for any μ such that $0 < \mu \leq 4$. In fact, a_0 increases as μ decreases. Typically, we set $\alpha = 0.05$, so $b_0 = \alpha/2 = 0.025$. Using the sequence of hyperparameters $\{\gamma_t\}_{t \in \mathbb{N}}$ (where $\gamma_t = 0.0722 \log(t \vee 2)/(t \exp(\sqrt{\log t}))$) as suggested by (Javanmard & Montanari, 2018) for this normal means model, we compute that $\gamma_1 \approx 0.117$ when the number of total hypotheses $T = 10^5$. Therefore, we get $a_0/(b_0\gamma_1) \approx 7.52 > 1$. As long as the weight ω is less than 7.52 with probability 1 under the alternative hypotheses (for any μ such that $0 < \mu < 4$), the condition needed for Theorem 4 is satisfied.*

6   Online FDR Control Experiments using Context-weighted GAI

In this section, we propose a practical procedure for contextual online FDR control based on context-weighted GAI rules, and present numerical experiments to illustrate the performance gains achievable with this procedure. Remember that a context-weighted GAI rule is a contextual GAI rule with

$$\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t) = \alpha_t(R_1, \ldots, R_{t-1}) \cdot \omega(X_t; \theta),$$

where $\omega(X_t; \theta)$ is a parametric function (with parameter set $\theta \in \Theta$) that is used to model the weight function. Technically, we can use any parametric function here. In this paper, we choose a deep neural network (multilayer perceptron) for this modeling due to its expressive power, as noted in a recent batch FDR control result by (Xia et al., 2017).

Given this, a natural goal will be to maximize the number of empirical discoveries while controlling FDR. To do so, we are going to find $\theta \in \Theta$ that maximizes the number of discoveries (or discovery rate), while empirically controlling the FDR. Note that if the function $\alpha_t(R_1, \ldots, R_{t-1})$ is monotone (such as with LORD or LORD++) with respect to the R_i's, the function $\alpha_t(R_1, \ldots, R_{t-1}) \cdot \omega(X_t; \theta)$ is also monotone with respect to the R_i's. In particular, this means that under the assumptions on the p-values from Theorem 1, FDR is controlled.

Training the Network, Setting θ.   Algorithm 1 is used for training the network, with the goal of adaptively increasing the discovery rate over time. Given a stream $((P_t, X_t))_{t \in \mathbb{N}}$, the algorithm processes the stream in batches, in a single pass. Let $b \geq 1$ denote the batch size. Let θ_j be the parameter obtained before batch j is processed, thus θ_j is only based on all previous p-values and contextual features which are assumed to be independent of all future batches. For each batch, the algorithm fixes the parameters to compute the significance levels for hypothesis in that batch. Define, the Empirical Discovery Rate for batch j as follows:

$$\text{EDR}_j = \frac{\sum_{i=jb+1}^{(j+1)b} \mathbb{1}\{P_i \leq \alpha_i(X_i; \theta)\}}{b}.$$

Since the above function is not differentiable, we use the sigmoid function σ to approximate the indicator function, and define

$$\text{EDR}_j = \frac{\sum_{i=jb+1}^{(j+1)b} \sigma(\lambda(\alpha_i(X_i; \theta) - P_i))}{b}.$$

Here λ is a large positive hyperparameter. With this, the parameter set θ can now be optimized by using standard (accelerated) gradient methods in an online fashion. Note that we are only maximizing empirical discovery rate subject to empirical FDR control, and the training does not require any ground truth labels

on the hypothesis. In fact, the intuition behind Algorithm 1 is very similar to the *policy gradient descent* method popular in reinforcement learning, which aims to find the best policy that optimizes the rewards. In an online multiple testing problem, we can regard the number of discoveries as reward and parametric functions as policies.

Algorithm 1: Online FDR Control with a Context-Weighted GAI Procedure

Model the weight function as a multi-layer perceptron (MLP);

Input: A sequence of p-value, contextual feature vector pairs $((P_1, X_1), (P_2, X_2), \dots)$, a monotone GAI rule (such as LORD++) denoted by $\mathbb{G}$ with desired FDR control, batch size b, learning rate η

Output: Neural network model parameter set $\theta \in \Theta$

Randomly initialize the parameter set θ_0; batch index $j = 0$
repeat
    for $i = 1$ *to* b **do**
        Consider the pair (P_{jb+i}, X_{jb+i}) (ith hypothesis in the jth batch)
        Let $\tilde{\alpha} \leftarrow \alpha_{jb+i}(R_1, \dots, R_{jb+i-1})$ (computed as defined by the GAI rule $\mathbb{G}$)
        Accept/reject this hypothesis, while satisfying GAI $\mathbb{G}$, with significance level reweighed as
          $\tilde{\alpha} \cdot \omega(X_{jb+i}; \theta_j)$
    end
    Use the decisions in the jth batch to update the empirical discovery proportion (EDR$_j$)
    Compute the gradient with respect to the parameter set $\frac{\partial \text{EDR}_j}{\partial \theta}$
    Update the parameter set: $\theta_{j+1} \leftarrow \theta_j + \eta \frac{\partial \text{EDR}_j}{\partial \theta}$
    $j \leftarrow j + 1$
until *convergence or end-of-stream*;
Return θ_j

In all our experiments, we use a multilayer perceptron to model the weight function, which is constructed by 10 layers and 10 nodes with ReLU as the activation function in each layer, and exponential function of the output layer, since the weight has to be non-negative. In the following, we use context-weighted LORD++ (**CwLORD++**) to denote the testing rule obtained from Algorithm 1 by using LORD++ as the monotone GAI rule.

6.1 Experiments

In this section, we present results for numerical experiments with both synthetic and real data to compare the performance of our proposed CwLORD++ testing rule with the current state-of-the-art online testing rule LORD++ (Ramdas et al., 2017b). The synthetic data experiments are based on the normal means model, which is commonly used in hypothesis testing literature for comparing the performance of various testing procedures. Our real data experiments focus on a diabetes prediction problem and gene expression data analyses. The primary goal of this section is to show the increased power (under FDR control) in online multiple testing setup that can be obtained with our proposed contextual weighting in many real world problems. Experiments with the SAFFRON procedure are presented in Appendix C.

6.1.1 Synthetic Data Experiments

For the synthetic data experiments, we consider the hypotheses $\mathcal{H}(T) = (H_1, \dots, H_T)$ coming from the normal means model. The setup is as follows: for $t \in [T]$, under the null hypothesis, $H_t : \mu_t = 0$, versus under the alternative, $\mu_t = \mu(X_t)$ is a function of X_t. We observe test statistics $Z_t = \mu_t + \varepsilon_t$, where ε_t's are

independent standard normal random variables, and thus the two-sided p-values are $P_t = 2\Phi(-|Z_t|)$. For simplicity, we consider a linear function $\mu(X_t) = \langle \boldsymbol{\beta}, X_t \rangle$ for $\boldsymbol{\beta}$ unknown to the testing setup. We choose the dimension of the features X_t's as $d = 10$ in all following experiments.
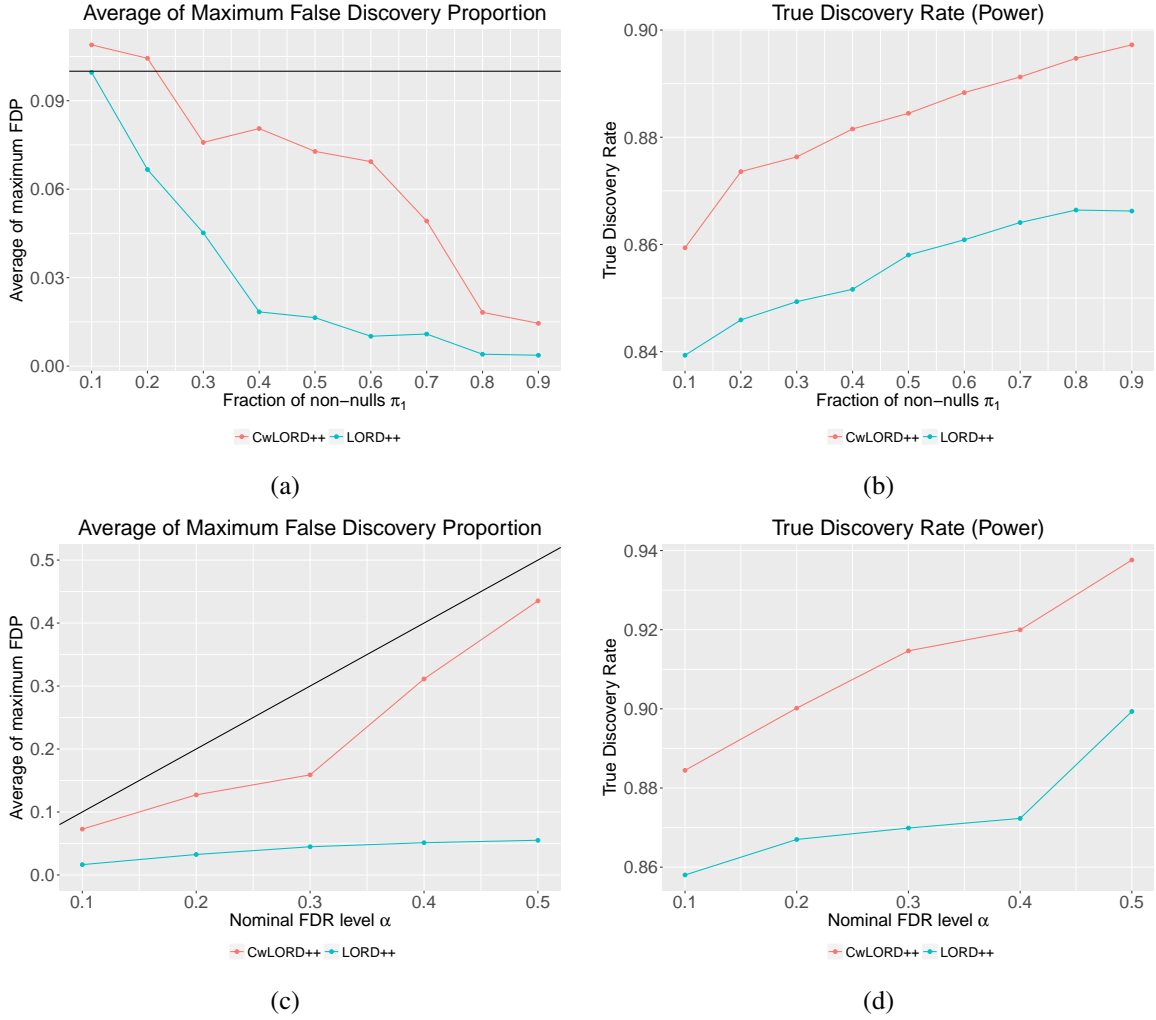

Figure 2: The top rows plots the average of max FDP and TDR (power) for our proposed CwLORD++ and LORD++ as we vary the fraction of non-nulls (π_1) under the normal means model. The nominal FDR control level $\alpha = 0.1$. The bottom row plots the same with varying nominal FDR levels. In this case, we set the fraction of non-nulls $\pi_1 = 0.5$. As mentioned in the text, the average of max FDP is an overestimate of FDR.

We set the total number of hypotheses as $T = 10^5$. We generate each d-dimensional vector X_t i.i.d. from $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = 2 \log T$. The choice of the σ^2 is to put the signals in a detectable (but not easy) region. This is because under the global null hypothesis where $Z_t \sim \mathcal{N}(0, 1)$ for all $t = 1, \ldots, T$, we have $\max_{t \in [T]} Z_t \sim \sqrt{2 \log T}$ with high probability. Here, $\boldsymbol{\beta}$ is a deterministic parameter vector of dimension $d = 10$, we generate the ith coordinate in $\boldsymbol{\beta}$ as $\beta_i \sim \text{Uniform}(-2, 2)$ and fix $\boldsymbol{\beta}$ throughout the following experiments. Let π_i denote the fraction of non-null hypotheses. For LORD++, we choose the sequence of hyperparameters $\{\gamma_t\}$ (where $\gamma_t = 0.0722 \log(t \vee 2)/(t \exp(\sqrt{\log t}))$) as suggested by (Javanmard & Montanari, 2018).

In Figure 2, we report the maximum FDP and the statistical power of the two compared procedures as

we vary the fraction of non-nulls π_1 and desired level α. The average is taken over 20 repeats.

In the first set of experiments (Figures 2(a) and 2(b)), we set $\alpha = 0.1$, and vary the fraction of non-nulls π_1 from 0.1 to 0.9. We can see that FDP of both rules (CwLORD++ and LORD++) are almost always under the set level $\alpha = 0.1$ and are decreasing with the increasing fraction of non-nulls. As expected, the power increases with increasing π_1, however the power of CwLORD++ uniformly dominates that of LORD++.

Note that we take the average of maximum FDP over 20 repeats, which is an estimate for $\mathbb{E}[\sup \text{FDP}]$. Due to the fact that $\mathbb{E}[\sup \text{FDP}] \geq \sup \mathbb{E}[\text{FDP}] = \sup \text{FDR}$, the reported average of maximum FDP is probably higher than the true maximum FDR. In Figure 2(a), we see that the average of maximum FDP is almost always controlled under the black line, which means the true maximum FDR should be even lower than that level. When π_1 is really small (like 0.1), the number of non-nulls is too sparse to make a high proportion of true discoveries, which leads to a higher average of maximum FDP that also have a higher variance in Figure 2(a).

In the second set of experiments (Figures 2(c) and 2(d)), we vary the nominal FDR level α from 0.1 to 0.5. The fraction of non-nulls is set as 0.5. Again we observe while both rules have FDR controlled under nominal level (the black line), and our proposed CwLORD++ is more powerful than the LORD++ with respect to the true discovery rate. On average, we notice about 3-5% improvement in the power with CwLORD++ when compared to LORD++.

6.1.2 Diabetes Prediction Problem

In this section, we apply our online multiple testing rules to a real-life application of diabetes prediction. Machine learning algorithms are now commonly used to construct predictive health scores for patients. In this particular problem, we want a test to identify patients that are at risk of developing diabetes. A high predicted risk score can trigger an intervention (such as medical follow-up, medical tests), which can be expensive and sometimes unnecessary, and therefore it is important to control the fraction of alerts that are false discoveries. That is, for each patient i, we form the null hypothesis H_i as the "patient will not develop diabetes" versus its alternative. The dataset was released as part of a Kaggle competition[5], which contains de-identified medical records of 9948 patients (labeled as $1, 2, \ldots$). For each patient, we have a response variable Y that indicates if the patient is diagnosed with Type 2 diabetes mellitus, along with patient's biographical information and details on medications, lab results, immunizations, allergies, and vital signs. In the following, we train a predictive score based on the available records, and then will apply our online multiple testing rule rules to control FDR on test set. Our overall methodology is similar to that used by Javanmard & Montanari (2018) in their FDR control experiments on this dataset. We proceed as follows. We construct the following features for each patient.

1. Biographical information: Age, height, weight, BMI (Body Mass Indicator), etc.

2. Medications: We construct TF-IDF vectors from the medication names.

3. Diagnosis information: We derive 20 categories from the ICD-9 codes and construct an one-hot encoded vector.

4. Physician specialty: We categorize the physician specialties and create features that represents how many times a patient visited certain specialist.

We regard the biographical information of patients as treated as contextual features. The choice of using biographical information as context is loosely based on the idea of *personalization* common in machine learning applications. Note that in theory, for our procedure, one could use other features too as context.

Second, we split the dataset into four parts **Train1**, comprising 40% of the data, **Train2**, 20% of the data, **Test1**, 20% of the data and **Test2**, 20% of the data. The **Train** sets are used for training a machine learning model (**Train1**) and for computing the null distribution of test statistics (**Train2**), which allows us

---

[5]http://www.kaggle.com/c/pf2012-diabetes

to compute the p-values in the **Test** sets. We first learn the neural network parameters in the CwLORD++ procedure in an online fashion by applying it to p-values in **Test1**, and then evaluate the performance of both LORD++ and CwLORD++ on **Test2**. This process is explained in more detail below.

We note that our experimental setup is not exactly identical to that of Javanmard & Montanari (2018), since we are using a slightly different set of features and data cleaning for the logistic regression model. We also split the data to four subsets instead of three as they did, which gives less training data for the predictive model. Our main focus, is to compare the power of LORD++ and CwLORD++, for a reasonable feature set and machine learning model.

Training Process.　We start by training a logistic model similar to (Javanmard & Montanari, 2018).[6] Let x_i denote the features of patient i. We use all the features to model the probability that patient does not have diabetes through a logistic regression model as

$$\Pr[Y_i = 0 \mid x = x_i] = \frac{1}{1 + \exp(\langle \beta, x_i \rangle)}.$$

The parameter β is estimated from the **Train1** set.

Construction of the p-values.　Let S_0 be the subset of patients in **Train2** set with labels as $Y = 0$, and let $n_0 = |S_0|$. For each $i \in S_0$, we compute its predictive score as $q_i = 1/(1 + \exp(\langle \beta, x_i \rangle))$. The empirical distribution of $\{q_i : i \in S_0\}$ serves as the null distribution of the test statistic, which allows for computation of the p-values. Explicitly, for each j in either **Test1** or **Test2** sets, we compute $q_j^{\text{Test}} = 1/(1 + \exp(\langle \beta, x_j \rangle))$, and construct the p-value P_j by

$$P_j = \frac{1}{n_0} \left| \{ i \in S_0 : q_i \le q_j^{\text{Test}} \} \right|.$$

Smaller p-value indicates that the patient has higher risk of developing diabetes. We use the p-values computed on the patients in **Test1** to train the weight function in CwLORD++, and the p-values on the patients in **Test2** to the compare performance of CwLORD++ and LORD++. Note that the training of the neural network does not utilize the labels of the hypothesis in the **Test1** set. Since the dataset does not have timestamps of hypotheses, we consider an ordering of hypotheses in the ascending order of corresponding p-values, and use this ordering for both LORD++ and CwLORD++. Note that, since the **Train** and **Test** sets are exchangeable, the null p-values will be uniform in expectation (and asymptotically uniform under mild conditions).

Online Hypothesis Testing Process and Results.　We set the desired FDR control level at $\alpha = 0.2$. The set of hyperparameters $\{\gamma_t\}$ is chosen as in the synthetic data experiments. For the patients in **Test1** set, we use their biographical information of patients as contextual features in the training process for CwLORD++ for learning the neural network parameters. We apply the LORD++ and CwLORD++ procedures to the p-values in the **Test2** set and compute the false discovery proportion and statistical power. Let T_2 be the set of patients in **Test2** set. Note that for a patient in **Test2** set, for both CwLORD++ and LORD++, the p-values are identically computed from all the features (including the patient's biographical information). This generates a sequence of p-values $(P_i)_{i \in T_2}$. Now, while LORD++ is applied to this p-value sequence directly, CwLORD++ is applied to the sequence of (P_i, X_i) where X_i is the biographical information of patient $i \in T_2$. For CwLORD++, the neural network parameters are fixed in this testing over the **Test2** set.

We repeat the whole process and average the results over 30 random splittings of the dataset. Table 1 presents our final result. We can use biographical information again as contextual features in training

---

[6]Even though the chosen logistic model is one of the best performing models on this dataset in the Kaggle competition, in this paper, we do not actively optimize the prediction model in the training process.

	FDR	Power
LORD++	0.147	0.384
CwLORD++	0.176	0.580

Table 1: Results from diabetes dataset with nominal FDR control level $\alpha = 0.2$.

CwLORD++ because the p-values under the null are uniformly distributed, no matter which features are used in logistic modeling. This guarantees that the p-values under the null are independent to any features, which is the only condition we need to have a FDR control, assuming the p-values themselves are mutually independent.

Notice that while FDR is under control for both procedures, the statistical power of CwLORD++ is substantially more (about 51%) than LORD++. This improvement illustrates the benefits of using contextual features for improving the power with FDR control in a typical machine learning setup. A possible reason for the observed increase in power with CwLORD++ is that in addition to using the labeled data in the Train1 set for training a supervised model, CwLORD++ uses in an unsupervised way (i.e., without considering labels) some features of the data in the **Test1** set in its online training process with the intent of maximizing discoveries.

One could also see the effect of training LORD++ on larger dataset. For example, previously we trained LORD++ only on **Train1** set (40% of the data) and we completely ignored the **Test1** set for LORD++. Suppose, we instead train logistic model for LORD++ (but not for CwLORD++) on the union of **Train1** and **Test1** set to make the baseline method stronger, with the same splits of the datasets. With this change, LORD++ has FDR at 0.143 and power at 0.427. While this is not completely a fair comparison for CwLORD++ as we have now used more labeled data in training the logistic model for LORD++ than CwLORD++, we observe that CwLORD++ still beats this stronger baseline with over 35% power increase.
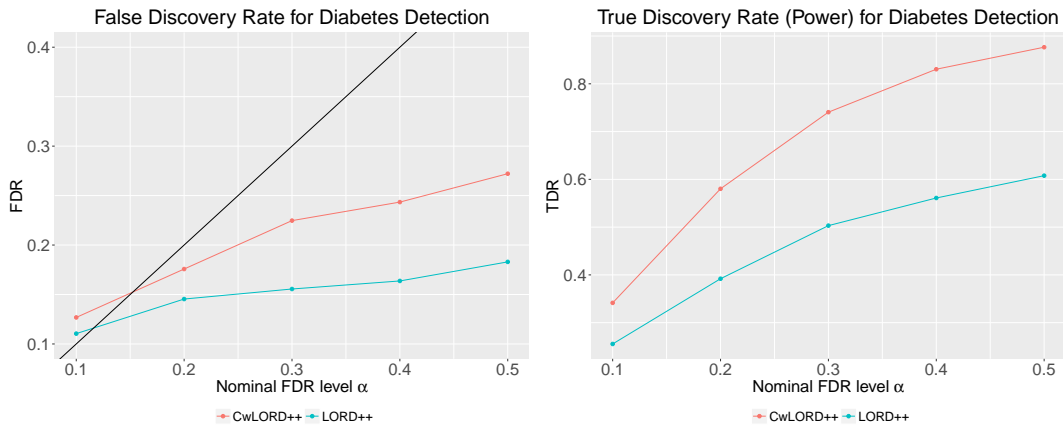

Figure 3: FDR and TDR results on diabetes dataset as we vary the nominal FDR level α. Note that the power of CwLORD++ uniformly dominates that of LORD++, with an average improvement in power of about 44%.

In order to further probe some of these improvements, we repeated the experiment with different nominal FDR levels ranging from 0.1 to 0.5. The results (see Figure 3) demonstrate that our CwLORD++ procedure achieves more true discoveries than the LORD++ procedure while controlling FDR under the same level. The FDR is controlled exactly under the desired level starting around $\alpha \geq 0.15$, while it is close to the desired level even when α is as small as 0.1. This phenomenon can also be observed in (Javanmard & Montanari, 2018), where the FDR is 0.126 for LORD under the target level $\alpha = 0.1$. This is probably because both

the experiments here and in (Javanmard & Montanari, 2018) do not adjust for the dependency among the p-values, which violates the theoretical assumption behind the FDR control proof, and is more of a concern when target α level is small.

6.1.3 Gene Expression Data

Our final set of experiments are on gene expression datasets. In particular, we use the Airway RNA-Seq and GTEx datasets[7] as also studied by (Xia et al., 2017). For both experiments, we use the original ordering of hypotheses as provided in the datasets. Since we don't know the ground truth, we only report the empirical FDR and the empirical discovery rate number in the experiments.

In the Airway RNA-Seq data, the aim is to identify glucocorticoid responsive (GC) genes that modulate cytokine function in airway smooth muscle cells. The dataset contains $n = 33469$ genes. The p-values are obtained in regular two-sample differential analysis of gene expression levels. Log counts of each gene serves as the contextual feature in this case. Figure 4 reports the empirical FDR and the discovery number. We see that our CwLORD++ procedure make about 10% more discoveries than the LORD++ procedure.

In the GTEx study, the question is to quantify the expression Quantitative Trait Loci (eQTLs) in human tissues. In the eQTL analysis, the association of each pair of single nucleotide polymorphism (SNP) and nearby gene is tested. The p-value is computed under the null hypothesis that the SNP genotype is not correlated with the gene expression. The GTEx dataset contains 464,636 pairs of SNP-gene combination from chromosome 1 in a brain tissue (interior caudate). Besides the p-values from the correlation test, contextual features may affect whether a SNP is likely to be an eQTL, and thus we can discover more of the true eQTLs if we utilize them in tests. For the tests, we consider the three contextual features studied by (Xia et al., 2017): 1) the distance (GTEx-dist) between the SNP and the gene (measured in log base-pairs); 2) the average expression (GTEx-exp) of the gene across individuals (measured in log rpkm); and 3) the evolutionary conservation measured by the standard PhastCons scores (GTEx-PhastCons). We apply LORD++ to the p-values, and CwLORD++ to the p-value, contextual feature vector pairs. Figure 5 reports the empirical FDR and the discovery number. In GTEx 1D experiments, we use each contextual feature separately in CwLORD++, which increases the discovery number by 5.5% (using GTEx-dist), 2.6% (using GTEx-exp), 2% (using GTEx-PhastCons) compared to the LORD++ procedure. In GTEx 2D experiments, GTEx-dist and GTEx-exp features are used together in CwLORD++ as a two-dimensional vector, and here CwLORD++ procedure make about 6.1% more discoveries than the LORD++ procedure. In GTEx 3D experiments, CwLORD++ uses all the three available features and makes about 6.6% more discoveries than LORD++. These results indicate that additional contextual information could be helpful in making more discoveries.

---

[7]Datasets are at: https://www.dropbox.com/sh/wtp58wd60980d6b/AAA4wA60ykP-fDfS5BNsNkiGa?dl=0.
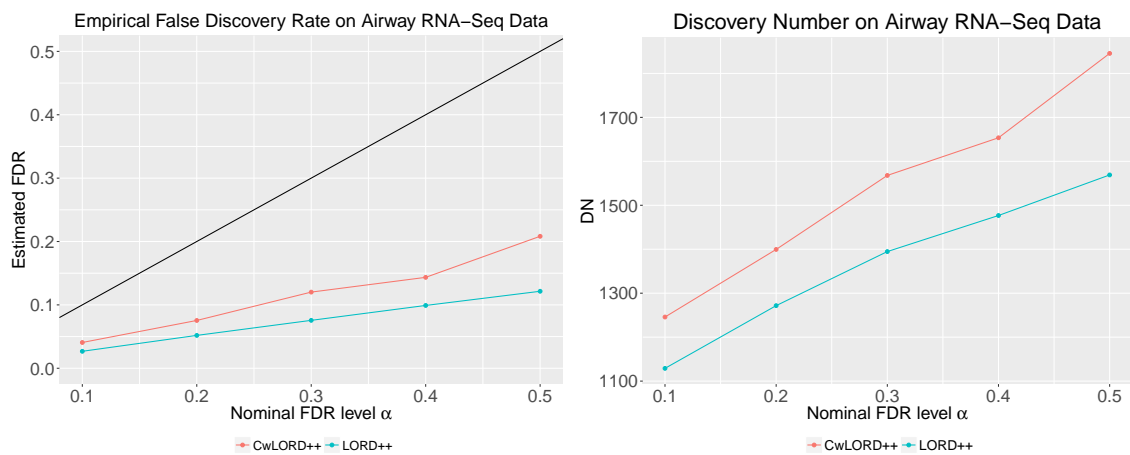
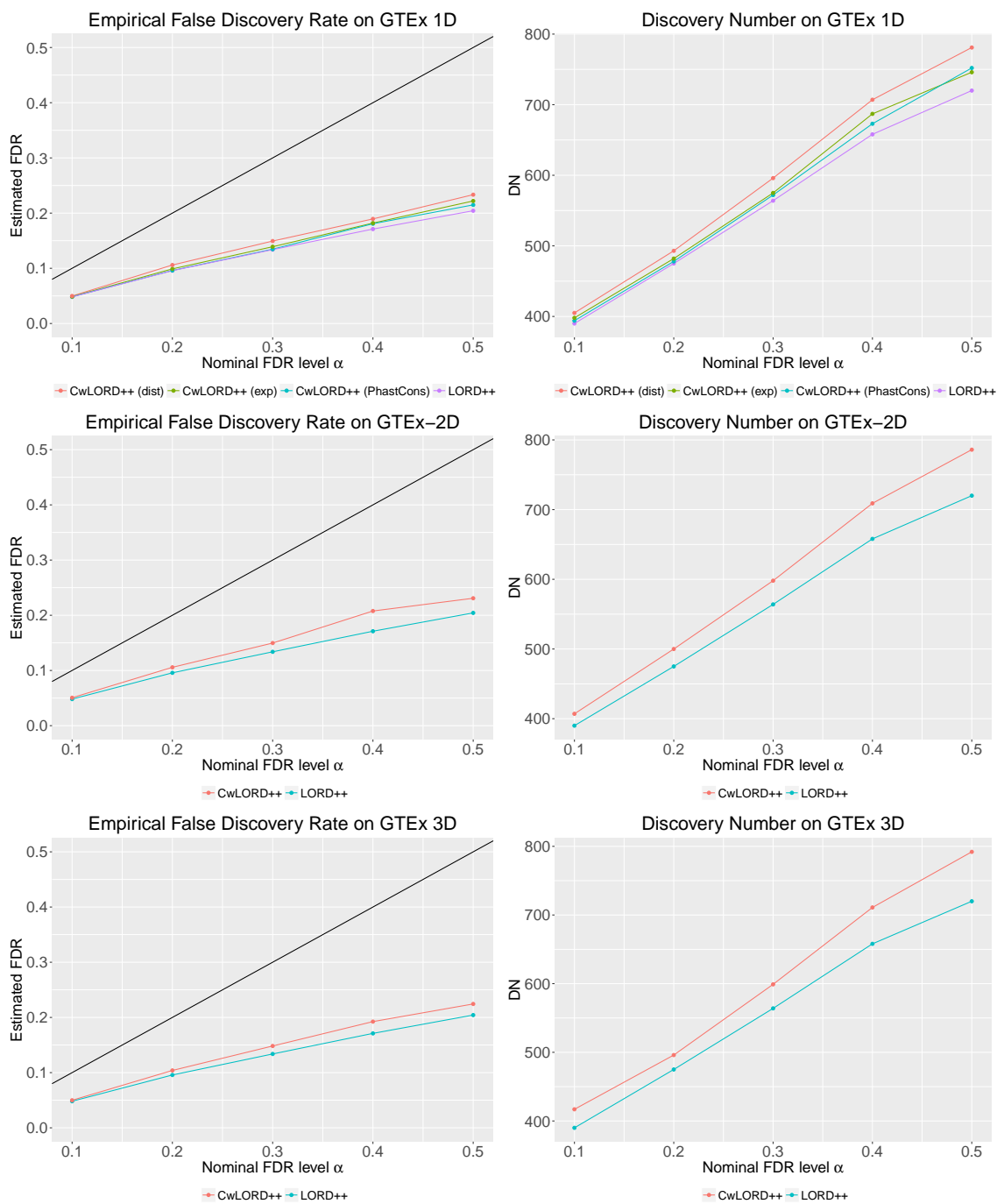Figure 4: Results on Airway RNA-Seq dataset.

Figure 5: Results on GTEx dataset.

7 Conclusion

In this paper, we introduced a new, rich class of online testing rules which incorporates the available contextual information in the testing process, and can control both the online FDR and mFDR under some standard assumptions. We then focused on a subclass of these rules, based on weighting the significance levels using contextual features, to derive a practical algorithm that learns a parametric weight function in an online fashion to maximize the number of empirical discoveries. We also theoretically proved that under standard informative-weighting assumptions, our procedure can achieve higher statistical power compared to a popular online FDR control procedure, while still controlling the FDR at the same level. Our experiments on both synthetic and real datasets demonstrates the superior performance of our procedure compared to other state-of-the-art online multiple testing procedures.

Acknowledgements

References

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

Benjamini, Y., & Hochberg, Y. (1995). 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.

Benjamini, Y., & Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, *24*(3), 407–418.

Carothers, N. L. (2000). *Real analysis*. Cambridge University Press.

Cox, D. R., Cox, D. R., Cox, D. R., & Cox, D. R. (1967). *Renewal theory*, vol. 1. Methuen London.

Dickhaus, T. (2014). *Simultaneous statistical inference*. Springer.

Dobriban, E. (2016). A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334*.

Dobriban, E., Fortney, K., Kim, S. K., & Owen, A. B. (2015). Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika*, *102*(4), 753–766.

Dudoit, S., & van der Laan, M. J. (2007). *Multiple testing procedures with applications to genomics*. Springer Science & Business Media.

Foster, D. P., & Stine, R. A. (2008). 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.

Foygel-Barber, R., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085.

Genovese, C. R., Roeder, K., & Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, *93*(3), 509–524.

G'Sell, M. G., Wager, S., Chouldechova, A., & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(2), 423–444.

Hu, J. X., Zhao, H., & Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, *105*(491), 1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, *13*(7), 577.

Javanmard, A., & Montanari, A. (2015). On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*.

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

Lei, L., & Fithian, W. (2016). Power of ordered hypothesis testing. *arXiv preprint arXiv:1606.01969*.

Lei, L., & Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(4), 649–679.

Li, A., & Barber, R. F. (2016a). Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, (just-accepted), 1–38.

Li, A., & Barber, R. F. (2016b). Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*.

Ramdas, A., Barber, R. F., Wainwright, M. J., & Jordan, M. I. (2017a). A unified treatment of multiple testing with prior knowledge using the p-filter. *arXiv preprint arXiv:1703.06222*.

Ramdas, A., Yang, F., Wainwright, M. J., & Jordan, M. I. (2017b). Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*, (pp. 5650–5659).

Ramdas, A., Zrnic, T., Wainwright, M., & Jordan, M. (2018). Saffron: an adaptive algorithm for online control of the false discovery rate. *arXiv preprint arXiv:1802.09098*.

Roquain, E. (2011). Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, *152*(2), 3–38.

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

Xia, F., Zhang, M. J., Zou, J. Y., & Tse, D. (2017). Neuralfdr: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, (pp. 1541–1550).

A Missing Details from Section 3

Identifiability of $f_1(p \mid X)$. We present a simple example from Lei & Fithian (2018) that illustrates why $f_1(p \mid X)$ (distribution of p under the alternate) is not identifiable. Consider the following mixture model:

$$H_t \mid X_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$

$$P_t \mid H_t, X_t = \begin{cases} \text{Uniform(0,1)} & \text{if } H_t = 0, \\ f_1(p \mid X_t) & \text{if } H_t = 1. \end{cases}$$

Now consider the conditional mixture density $f(p \mid X) = (1 - \pi_1) + \pi_1 f_1(p \mid X)$. Note that the H_t's are not observed. Thus, while f is identifiable from the data, π_1 and f_1 are not: for example, $\pi_1 = 0.5$, $f_1(p \mid X) = 2(1 - p)$ and $\pi_1 = 1$, $f_1(p \mid X) = 1.5 - p$ result in exactly the same mixture density $f(p \mid X)$.

Lemma 2 (Lemma 1 Restated). *Let $g : \{0, 1\}^T \to \mathbb{R}$ be any coordinatewise non-decreasing function such that $g(\mathbf{R}) > 0$ for any vector $\mathbf{R} \neq (0, \ldots, 0)$. Then for any index $t \leq T$ such that $t \in \mathcal{H}^0$, we have*

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(R_1, \ldots, R_T) \vee 1} \Big| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right] \leq \mathbb{E}\left[\frac{\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_T)}{g(R_1, \ldots, R_T) \vee 1} \Big| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right].$$

Proof. Let $\mathbf{P} = (P_1, \ldots, P_T)$ be the sequence of p-values, and $\mathbf{X} = (X_1, \ldots, X_T)$ be the sequence of the contextual feature vectors until sometime T. We define a "leave-one-out" vector of p-value as $\widetilde{\mathbf{P}}^{-t} = (\widetilde{P}_1, \ldots, \widetilde{P}_T)$, which was obtained from $\mathbf{P}$ by setting $P_t = 0$, i.e.,

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

Let $\mathbf{R} = (R_1, \ldots, R_T)$ be the sequence of decisions on the input $\mathbf{P}$ and $\mathbf{X}$, and $\widetilde{\mathbf{R}}^{-t} = (\widetilde{R}_1, \ldots, \widetilde{R}_T)$ be the sequence of decisions by applying the same rule on the input $\mathbf{P}^{-t}$ and $\mathbf{X}$. Note here we just set one p-value as zero but are not changing the contextual feature vectors.

By the construction of p-values, we have that $R_i = \widetilde{R}_i$ for $i < t$, and hence

$$\alpha_i(R_1, \ldots, R_{i-1}, X_1, \ldots, X_i) = \alpha_i(\widetilde{R}_1, \ldots, \widetilde{R}_{i-1}, X_1, \ldots, X_i), \quad \text{for all } i \leq t.$$

We also know that $\widetilde{R}_t = 1$ always holds due to the fact $\widetilde{P}_t = 0 \leq \alpha_t$. Therefore, if the event $\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}$ occurs, we have $R_t = \widetilde{R}_t$ and thus $\mathbf{R} = \widetilde{\mathbf{R}}^{-t}$.

From the above arguments, we conclude that

$$\frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\mathbf{R}) \vee 1} = \frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1}.$$

Due to the fact that $t \in \mathcal{H}^0$ ($H_t = 0$), P_t is independent to all contextual features $\mathbf{X}$ by assumption (as P_i's and X_i's are independent under the null), which gives that P_t is independent of $\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$. And since P_t is independent of $\widetilde{\mathbf{R}}^{-t}$, we have,

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\mathbf{R}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right]$$

$$= \mathbb{E}\left[\frac{\mathbb{1}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right]$$

$$\leq \mathbb{E}\left[\frac{\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right] \tag{22}$$

$$\leq \mathbb{E}\left[\frac{\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)}{g(\mathbf{R}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t) \right] \tag{23}$$

where inequality (22) follows by taking expectation with respect to P_t and using conditional super-uniformity (12), and inequality (23) is derived by the following observation.

Since $\widetilde{P}_t = 0 \leq \alpha_t$, we have $\widetilde{R}_t = 1 \geq R_t$. Due to the monotonicity of the significance levels, we have

$$\alpha_i(\widetilde{R}_1, \ldots, \widetilde{R}_{i-1}, X_1, \ldots, X_i) \geq \alpha_i(R_1, \ldots, R_{i-1}, X_1, \ldots, X_i), \quad \text{for all } i > t,$$

ensuring $\widetilde{R}_i \geq R_i$ for all i, and thus $g(\widetilde{\mathbf{R}}^{-t}) \geq g(\mathbf{R})$ by the non-decreasing assumption on the function g. $\qquad\square$

Theorem 5 (Theorem 1 Restated). *Consider a sequence of $((P_t, X_t))_{t \in \mathbb{N}}$ of p-values and contextual features. If the p-values P_t's are independent, and additionally P_t are independent of all $(X_t)_{t \in \mathbb{N}}$ under the null (whenever $H_t = 0$), then for any monotone contextual generalized alpha-investing rule (i.e., satisfying conditions (7), (8), (9), (10), and (11)), we have online FDR control,*

$$\sup_T \text{FDR}(T) \leq \alpha.$$

Proof. Note that the number of false discoveries is $V(T) = \sum_{t=1}^T R_t \mathbb{1}\{t \in \mathcal{H}^0\}$ and the amount of wealth is $W(T) = w_0 + \sum_{t=1}^T (-\phi_t + R_t \psi_t)$.

We can derive the following expression by using the tower property of conditional expectation

$$\mathbb{E}\left[\frac{V(T) + W(T)}{R(T) \vee 1}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\frac{R_t \mathbb{1}\{t \in \mathcal{H}^0\} + \frac{w_0}{T} - \phi_t + R_t \psi_t}{R(T) \vee 1}\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right] \qquad (24)$$

We split the analysis in two cases based on whether $H_t = 0$ or $H_t = 1$.

- Case 1: Suppose that $t \in \mathcal{H}^0$. By applying Lemma 1, we have

$$\mathbb{E}\left[\frac{R_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] = \mathbb{E}\left[\frac{\mathbb{1}\{P_t \leq \alpha_t\}}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]$$

$$\leq \mathbb{E}\left[\frac{\alpha_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]. \qquad (25)$$

Since $\psi_t \leq \frac{\phi_t}{\alpha_t} + b_t - 1$, we further obtain

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right] \leq \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\frac{\phi_t}{\alpha_t} + b_t) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t b_t + \frac{\phi_t}{\alpha_t}(R_t - \alpha_t)}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right]$$

$$\leq \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t b_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right],$$

where the last inequality follows by applying (25).

- Case 2: Suppose that $t \notin \mathcal{H}^0$. Using the fact that $\psi_t \leq \phi_t + b_t$, we have

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right] \leq \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\phi_t + b_t) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t b_t + (R_t - 1)\phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right]$$

$$\leq \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t b_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right].$$

Combining the bound on $\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right]$ from both cases in (24) and using the definition of b_t, we obtain that,

$$\mathbb{E}\left[\frac{V(T) + W(T)}{R(T) \vee 1}\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\frac{w_0}{T} + R_t b_t}{R(T) \vee 1}\right] = \mathbb{E}\left[\frac{w_0 + \sum_{t=1}^{T} R_t b_t}{R(T) \vee 1}\right]$$

$$\leq \mathbb{E}\left[\frac{w_0 + \sum_{t=1}^{T} R_t \alpha - w_0 \mathbb{1}\{T \geq \rho_1\}}{R(T) \vee 1}\right] = \mathbb{E}\left[\frac{w_0 + \alpha R(T) - w_0 \mathbb{1}\{T \geq \rho_1\}}{R(T) \vee 1}\right] \leq \alpha.$$

This concludes the proof of the theorem. □

Theorem 6 (Theorem 2 Restated). *Consider a sequence of $((P_t, X_t))_{t \in \mathbb{N}}$ of p-values and contextual features. If the p-values P_t's are conditionally super-uniform distributed (as in (??)) , then for any contextual generalized alpha-investing rule (i.e., satisfying conditions (7), (8), (9), and (10)), we have online mFDR control,*

$$\sup_{T \in \mathbb{N}} \text{mFDR}(T) \leq \alpha.$$

Proof. The conditional super-uniformity implies that under null

$$\mathbb{E}\left[R_t \big| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \leq \alpha_t.$$

Now using a proof technique similar to Theorem 5, for any $T \in \mathbb{N}$, we get

$$
\begin{aligned}
\mathbb{E}[V(T)] &\leq \mathbb{E}[V(T) + W(T)] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[R_t \mathbb{1}\{t \in \mathcal{H}^0\} + \frac{w_0}{T} - \phi_t + R_t \psi_t\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t \big| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{w_0}{T} + R_t b_t \big| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]\right] \\
&= \mathbb{E}\left[w_0 + \sum_{t=1}^{T} R_t b_t\right] = \mathbb{E}\left[w_0 + \alpha R(T) - w_0 \mathbb{1}\{T \geq \rho_1\}\right] \\
&\leq \alpha \mathbb{E}\left[R(T) \vee 1\right],
\end{aligned}
$$

where for the second inequality we used an analysis similar to that used in the first case in the proof of Theorem 5. Therefore, for any $T \in \mathbb{N}$,

$$\text{mFDR(T)} = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T) \vee 1]} \leq \alpha.$$

This concludes the proof of the theorem. $\qquad\square$

B   Missing Details from Section 5

Proposition 3 (Proposition 1 Restated). *Suppose that the weight distribution satisfies the informative-weighting assumption in (16). Suppose that p-values P_t's are independent, and are conditionally independent of the weights ω_t's given H_t's. Then the weighted LORD++ rule can control the online FDR at any given level α, i.e.,*

$$\sup_{T \in \mathbb{N}} \text{FDR}(T) \leq \alpha.$$

Proof. We start with a frequently used estimator of FDR that is defined as:

$$\widehat{\text{FDP}}(T) := \frac{\sum_{t=1}^{T} \alpha_t}{R(T) \vee 1}.$$

As established in Section 4 in Ramdas et al. (2017b), LORD++ applied to any sequence of p-values will ensure that $\sup_T \widehat{\text{FDP}}(T) \le \alpha$. We apply LORD++ with the sequence of p-values defined as $\boldsymbol{P}' = \left(\frac{P_1}{\omega_1}, \frac{P_2}{\omega_2}, \frac{P_3}{\omega_3} \dots\right)$. Let $P'_t = P_t / \omega_t$ for any $t \in \mathbb{N}$. Then it follows that,

$$\sup_{T \in \mathbb{N}} \widehat{\text{FDP}}(T) = \sup_{T \in \mathbb{N}} \frac{\sum_{t=1}^T \alpha_t}{R(T) \vee 1} = \sup_{T \in \mathbb{N}} \frac{\sum_{t=1}^T \alpha_t}{\left(\sum_{t=1}^T \mathbb{1}\{P'_t \le \alpha_t\}\right) \vee 1} \le \alpha. \tag{26}$$

We denote the sigma-field of decisions based on the weighted p-values $\boldsymbol{P}'$ till time t as $\mathcal{C}^t = \sigma(R_1, \dots, R_t)$. By using the "leave-one-out" method used in Theorem 5, the FDR of the weighted LORD++ at any time T can be written as,

$$\begin{aligned}
\text{FDR}(T) &= \mathbb{E}\left[\frac{\sum_{t=1}^T \mathbb{1}\{t \in \mathcal{H}^0 : P'_t \le \alpha_t\}}{\left(\sum_{t=1}^T \mathbb{1}\{P'_t \le \alpha_t\}\right) \vee 1}\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\frac{\mathbb{1}\{t \in \mathcal{H}^0 : \frac{P_t}{\omega_t} \le \alpha_t\}}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{t \in \mathcal{H}^0 : \frac{P_t}{\omega_t} \le \alpha_t\}}{R(T) \vee 1}\Big|\mathcal{C}^{t-1}\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{t \in \mathcal{H}^0 : \frac{P_t}{\omega_t} \le \alpha_t\}}{R^{-t}(T) \vee 1}\Big|\mathcal{C}^{t-1}\right]\right],
\end{aligned}$$

where $R^{-t}(T) = \sum_{i=1}^T \mathbb{1}\{P_i / \omega_i \le \alpha_i\}$ is obtained by setting $P_t = 0$, while keeping all ω_t's unchanged. The last equality holds due to the fact that $R^{-t}(T) = R(T)$ given the event $\{P_t / \omega \le \alpha_t\}$.

Since $\alpha_t \in \mathcal{C}^{t-1}$, and P_t, ω_t are independent of $R^{-t}(T)$ and $\mathcal{C}^{t-1}$, we can take the expectation of the numerator inside the brackets and obtain that

$$\begin{aligned}
\Pr[P_t / \omega_t \le \alpha_t \mid \mathcal{C}^{t-1}, H_t = 0] &= \int \Pr[P_t / \omega_t \le \alpha_t \mid \mathcal{C}^{t-1}, \omega_t = w, H_t = 0] \, dQ(w \mid H_t = 0) \\
&= \int w \alpha_t dQ(w \mid H_t = 0) \\
&= u_0 \alpha_t,
\end{aligned}$$

where $u_0 = \mathbb{E}[\omega \mid H_t = 0]$. Plugging this in the bound on $\text{FDR}(T)$ from above gives,

$$\begin{aligned}
\text{FDR}(T) &= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{u_0 \alpha_t}{R^{-t}(T) \vee 1}\Big|\mathcal{C}^{t-1}\right]\right] \\
&\le \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_t}{R^{-t}(T) \vee 1}\Big|\mathcal{C}^{t-1}\right]\right] \tag{27} \\
&\le \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_t}{R(T) \vee 1}\Big|\mathcal{C}^{t-1}\right]\right] \tag{28} \\
&= \mathbb{E}\left[\frac{\sum_{t=1}^T \alpha_t}{R(T) \vee 1}\right] = \mathbb{E}\left[\widehat{\text{FDP}}(T)\right] \le \alpha,
\end{aligned}$$

where inequality (27) is due to the assumption that $u_0 < 1$, (28) follows by the fact that $R^{-t}(T) \ge R(T)$ due to monotonicity of LORD++, and the last equality is based on (26). $\qquad\square$

Theorem 7 (Theorem 3 Restated). *Let $D(a) = \Pr[P / \omega \le a]$ be the above marginal distribution of weighted p-values. Then, the average power of weighted LORD++ rule is almost surely bounded as follows:*

$$\liminf_{T \to \infty} \text{TDR}(T) \ge \left(\sum_{m=1}^\infty \prod_{j=1}^m (1 - D(b_0 \gamma_j))\right)^{-1}.$$

Proof. Since we are interested in lower bounds, we consider a version of LORD (as also considered in (Javanmard & Montanari, 2018)) which that is based on the following rule,

$$\textbf{LORD:}W(0) = w_0 = b_0 = \alpha/2, \quad \phi_t = \alpha_t = b_0\gamma_{t-\tau_t}, \quad \psi_t = b_0.$$

Note that since $b_t = \alpha - w_0\mathbb{1}\{\rho_1 > t-1\} > b_0$ in LORD++, the test level in LORD++ is at least as large to the test level in LORD. Therefore, for any p-value sequence the power of the LORD from is also a lower bound on the power of LORD++. In the rest of this proof, we focus on LORD for the weighted p-value sequence $\{P_1/\omega_1, P_2/\omega_2, \dots\}$. The bound established below is in fact *tight* for LORD under this p-value sequence.

Denote by ρ_i as the time of the ith discovery (rejection), with $\rho_0 = 0$, and $\Delta_i = \rho_i - \rho_{i-1}$ as the ith time interval between the $(i-1)$st and ith discoveries. Let $r_i := \mathbb{1}\{\rho_i \in \mathcal{H}^1\}$ be the reward associated with inter-discovery Δ_i. Since the weighted p-values are i.i.d. it can be seen that the times between successive discoveries are i.i.d. according to the testing procedure LORD, and the process $R(T) = \sum_{l=1}^{T} R_l$ is a *renewal process* (Cox et al., 1967). In fact, for each i, we have

$$\Pr[\Delta_i \geq m] = \Pr[\cap_{l=\rho_{i-1}}^{\rho_{i-1}+m}\{P_l/\omega_l > \alpha_l\}] = \prod_{l=\rho_{i-1}}^{\rho_{i-1}+m}(1 - D(\alpha_l)) = \prod_{l=\rho_{i-1}}^{\rho_{i-1}+m}(1 - D(b_0\gamma_{l-\rho_{i-1}})) = \prod_{l=1}^{m}(1 - D(b_0\gamma_l)).$$

The above expression is same for every i. Therefore,

$$\mathbb{E}[\Delta_i] = \sum_{m=1}^{\infty}\Pr[\Delta_i \geq m] = \sum_{m=1}^{\infty}\prod_{l=1}^{m}(1 - D(b_0\gamma_l)).$$

Applying the strong law of large numbers for renewal-reward processes (Cox et al., 1967), we obtain that the following statement holds almost surely,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{R(T)} r_i = \frac{\mathbb{E}(r_i)}{\mathbb{E}(\Delta_1)} = \pi_1\big(\sum_{m=1}^{\infty}\prod_{l=1}^{m}(1 - D(b_0\gamma_l))\big)^{-1}.$$

Let $|\mathcal{H}^1(T)|$ be the number of true alternatives till time T. Since $\lim_{T\to\infty}|\mathcal{H}^1(T)|/T = \pi_1$ almost surely, we have

$$\lim_{T\to\infty}\frac{1}{|\mathcal{H}^1(T)|}\sum_{i\in\mathcal{H}^1(T)} R_i = \lim_{T\to\infty}\frac{1}{|\mathcal{H}^1(T)|}\sum_{i=1}^{R(T)} r_i = \big(\sum_{m=1}^{\infty}\prod_{l=1}^{m}(1 - D(b_0\gamma_l))\big)^{-1}.$$

Now by using the definition of $\text{TDP}(T)$, almost surely, we have that for any weighted LORD++,

$$\liminf_{T\to\infty}\text{TDP}(T) \geq \big(\sum_{m=1}^{\infty}\prod_{j=1}^{m}(1 - D(b_0\gamma_j))\big)^{-1}.$$

As discussed above, this bound translates into a lower bound for weighted LORD++. Furthermore, by using the Fatou's lemma (Carothers, 2000), we can extend the same result for $\text{TDR}(T)$ almost surely,

$$\liminf_{T\to\infty}\text{TDR}(T) = \liminf_{T\to\infty}\mathbb{E}[\text{TDP}(T)] \geq \mathbb{E}[\liminf_{T\to\infty}\text{TDP}(T)] \geq \big(\sum_{m=1}^{\infty}\prod_{j=1}^{m}(1 - D(b_0\gamma_j))\big)^{-1}.$$

$\square$

Theorem 8 (Theorem 4 Restated). *Suppose that the parameters in LORD (19) satisfy $b_0\gamma_1 < a_0$, and the weight distribution satisfies $\Pr[\omega < a_0/(b_0\gamma_1) \mid H_t = 1] = 1$ for every $t \in \mathbb{N}$ and the informative-weighting assumption in (16). Then, the average power of weighted LORD is greater than equal to that of LORD almost surely.*

Proof. We compare the average power bound of weighted LORD and LORD. It is equivalent to comparing $D(a)$ and $G(a)$ for $a = b_0\gamma_l$, for $l = 1, \ldots, \infty$. Since $u = (1-\pi_1)u_0+\pi_1 u_1 = 1$, we have $(1-\pi_1)u_0 = 1-\pi_1 u_1$. This means that

$$D(a) - G(a) = (1 - \pi_1)u_0 a + \pi_1 \int F(aw)\, dQ_1(w) - (1 - \pi_1)a - \pi_1 F(a)$$

$$= (1 - \pi_1)(u_0 - 1)a + \pi_1 \left(\int F(aw)\, dQ_1(w) - F(a) \right)$$

$$= \pi_1(1 - u_1)a + \pi_1 \left(\int F(aw)\, dQ_1(w) - F(a) \right).$$

So we just need to compare $(\mu_1 - 1)a$ and $\int F(aw)\, dQ_1(w) - F(a)$, for any $a = b_0\gamma_l$, for $l = 1, \ldots, \infty$. Due to the fact that $\{\gamma_l\}$ is a non-increasing sequence, we have $a = b_0\gamma_l \leq b_0\gamma_1$. Since $b_0\gamma_1 < a_0$ and $\Pr[\omega < a_0/(b_0\gamma_1) \mid H = 1] = 1$ by assumption, then $\Pr[\max(a, aw) < a_0 \mid H = 1] = 1$.

For any fixed $a = b_0\gamma_l > 0$, we have

$$\frac{\int F(aw)\, dQ_1(w) - F(a)}{a} = \int \frac{F(aw) - F(a)}{a}\, dQ_1(w)$$

$$= \int \frac{F(aw) - F(a)}{(w - 1)a}(w - 1)\, dQ_1(w)$$

$$= \int f(\xi)(w - 1)\, dQ_1(w) \tag{29}$$

$$\geq \int (w - 1)\, dQ_1(w) \tag{30}$$

$$= \mathbb{E}[W \mid H = 1] - 1 = u_1 - 1,$$

for some $\xi \in (\min(a, aw), \max(a, aw))$. Note we assume Q_1 is a continuous distribution, so $\Pr[w = 1 \mid H = 1] = 0$. The equality (29) is achieved by applying the Intermediate Value Theorem, and the inequality (30) is obtained by the fact that $\Pr[\xi < a_0 \mid H = 1] = 1$, i.e., $\Pr[f(\xi) > 1 \mid H = 1] = 1$.

Therefore, we prove that $\int F(aw)\, dQ_1(w) - F(a) \geq u_1 - 1$, which implies that $D(a) \geq G(a)$ for $a = b_0\gamma_l$, for $l = 1, \ldots, \infty$. $\qquad\square$

C SAFFRON Procedure

Let us start with a quick introduction to the SAFFRON procedure proposed by Ramdas et al. (2018). Since SAFFRON can be viewed as an online analogue of the famous offline Storey-BH adaptive procedure (Storey, 2002), we start a description of the Storey-BH procedure.

In the offline setting where p-values are all available, the rejection rule is to reject all p-values below some threshold s, meaning that $\mathcal{R}(s) = \{i \mid P_i \leq s\}$. Thus an oracle estimate for FDP is given by

$$\mathrm{FDP}^*(s) := \frac{|\mathcal{H}^0| \cdot s}{|\mathcal{R}(s)| \vee 1}.$$

The world oracle means that FDP* cannot be calculated, since $\mathcal{H}^0$ is unknown. The BH method overestimates FDP$^*(s)$ by the empirically computable quantity

$$\widehat{\mathrm{FDP}}_{\mathrm{BH}}(s) := \frac{n \cdot s}{|\mathcal{R}(s)| \vee 1},$$

and chooses the threshold $\hat{s}_{\mathrm{BH}}(s) = \max\{s : \widehat{\mathrm{FDP}}_{\mathrm{BH}}(s) \leq \alpha\}$. However, Storey (2002) noted that the estimate of $\widehat{\mathrm{FDP}}_{\mathrm{BH}}(s)$ is conservative, and thus proposed a different estimate (referred to as Storey-BH) as

$$\widehat{\mathrm{FDP}}_{\mathrm{St\text{-}BH}}(s) := \frac{n \cdot s \cdot \hat{\pi}_0}{|\mathcal{R}(s)| \vee 1},$$

where the fraction of nulls $\hat{\pi}_0$ is estimated by

$$\hat{\pi}_0 := \frac{1 + \sum_{i=1}^n \mathbb{1}\{P_i > \lambda\}}{n(1 - \lambda)},$$

with a well-chosen λ. There is a bias-variance trade-off in the choice of λ. When λ grows larger, the bias of $\hat{\pi}_0$ grows smaller while the variance becomes larger. Through numerical simulations Storey (2002) demonstrated that there could be an increase in power (over the BH procedure) with this adaptivity.

Similarly, in the online setting, the oracle FDP estimate now is

$$\text{FDP}^*(T) := \frac{\sum_{t \in [T], t \in \mathcal{H}^0} \alpha_t}{R(T) \vee 1}$$

The connection between SAFFRON and LORD/LORD++ is the same as that between Storey-BH and BH. Empirically, LORD/LORD++ overestimates the oracle $\text{FDP}^*(T)$ as

$$\widehat{\text{FDP}}_{\text{LORD}}(T) := \frac{\sum_{t \in [T]} \alpha_t}{R(T) \vee 1}$$

SAFFRON estimates the amount of alpha-wealth that was spent testing nulls so far, which is analogous to the proportion of nulls in the offline setting, and controls the following overestimate of oracle FDP,

$$\widehat{\text{FDP}}_{\text{SAFFRON}}(T) := \frac{\sum_{t \in [T]} \alpha_t \frac{\mathbb{1}\{P_t > \lambda_t\}}{(1 - \lambda_t)}}{R(T) \vee 1},$$

where $\{\lambda_t\}_{t=1}^\infty$ is predictable sequence of user-chosen parameters in interval $(0, 1)$. Note that when $\lambda_t = 0$, it recovers $\widehat{\text{FDR}}_{\text{LORD}}(T)$. Ramdas et al. (2018) proved that, under some constraints, SAFFRON can control online FDR at given level α by showing that

$$\text{FDR}(T) \leq \mathbb{E}[\widehat{\text{FDP}}_{\text{SAFFRON}}(T)].$$

However, when we extend SAFFRON to incorporate contextual features, due to the optimization problem that we solve in Algorithm 1, it seems difficult to ensure FDR control in practice. The possible reason here is analogous to overfitting in machine learning, due to the bias-variance tradeoff inherent in the SAFFRON estimator. When λ_t's is set large, the bias of the estimator becomes smaller but then the variance increases. And when we optimize an FDR estimator with large variance in the training process (meaning the precision of FDR estimator is poor), we will lose FDR control in the validation process, just like overfitting.

We still present some empirical evidence that contextual information could help with SAFFRON too, though one has to be careful about FDR control.

Experiments with SAFFRON. We consider exactly the same setting as with the synthetic data experiments in Section 6.1.1, and train a context-weighted SAFFRON (referred to as CwSAFFRON) in the same way as CwLORD++. With varying fraction of non-nulls, Figure 6 reports the maximum FDP and statistical power (TDP) of CwSAFFRON along with three other procedures, SAFFRON, CwLORD++, and LORD++. We observe that SAFFRON and CwSAFFRON have FDR greater than the nominal level of 0.1 when the fraction of non-nulls π_1 is small (less than 0.3), but is below the nominal level when π_1 gets larger. And the FDR of both LORD++ and CwLORD++ are generally much smaller than that of SAFFRON and CwSAFFRON. This is consistent with our discussion about SAFFRON in Section 2. On the other hand, the power of CwSAFFRON dominates that of SAFFRON for all π_1, and is larger than that of CwLORD++ when π_1

exceeds 0.3. As the fraction of non-nulls increases, CwSAFFRON achieves a faster increase in power than CwLORD++. A similar phenomenon can be also seen between SAFFRON and LORD++ (which was also noted by Ramdas et al. (2018)).
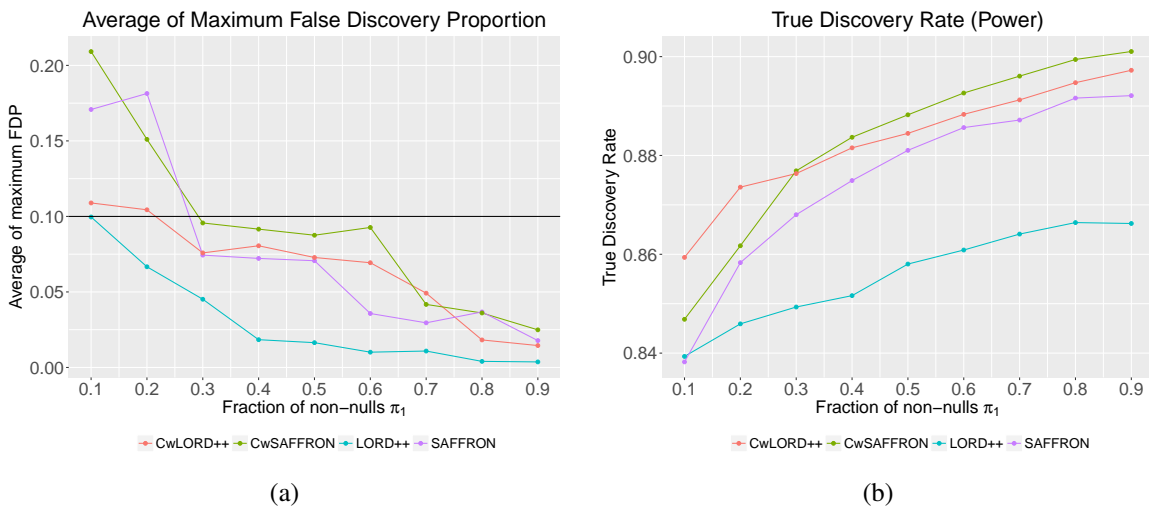

Figure 6: The average of max FDP and TDR (power) for our proposed CwSAFFRON, CwLORD++, along with SAFFRON and LORD++ with varying the fraction non-nulls (π_1) under the normal means model. The nominal FDR control level $\alpha = 0.1$. As mentioned in the description of the synthetic data experiments, the average of max FDP is an overestimate of FDR.