Contextual Online False Discovery Rate Control

Shiyun Chen
University of California San Diego

Shiva Prasad Kasiviswanathan
Amazon, Sunnyvale, USA

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 a setting where an ordered (possibly infinite) sequence of hypotheses arrives in a stream, 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. This model provides a general way of leveraging the side (contextual) information in the data to help maximize the number of discoveries while controlling the FDR.

We propose a new class of powerful online testing procedures, where the rejection thresholds are learned 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 the rejection thresholds, to derive a practical algorithm that learns a parametric weight function in an online fashion to gain more discoveries. We also theoretically prove that our proposed procedures, under some easily verifiable assumptions, would lead to an increase of statistical power over a popular online testing procedure proposed by (Javanmard and Montanari, 2018). Finally, we demonstrate the superior 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.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 Introduction

Multiple hypotheses testing - controlling overall error rates when performing multiple hypothesis tests - is a well-established area in statistics with applications in a variety of scientific disciplines (Dudoit and van der Laan, 2007; Dickhaus, 2014; 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, or is this gene mutation associated with certain trait?

Typically, each hypothesis is summarized to one p-value, and is rejected (or claimed as a non-null) if the p-value is below some significance level. The rejected hypotheses are called *discoveries*, and those that were true nulls 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, where we have an entire batch of hypotheses and the corresponding p-values, and (Benjamini and Hochberg, 1995) developed a standard procedure (called *BH procedure*) to control FDR below a preassigned level. 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 and Stine, 2008) (a 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 decisions before next hypothesis arrives, without access to the number of hypotheses in the stream or future p-values. 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 trials are conducted over time, or with publicly available datasets where new hypotheses are tested in an on-going fashion by different researchers.

Foster and Stine (2008) designed the first online alpha-investing procedures that use and earn alpha-wealth to control a modified variant of FDR (referred to as *mFDR*), which was later extended to a class of *generalized alpha-investing*

(GAI) rules by (Aharoni and Rosset, 2014). Javanmard and Montanari (2015, 2018) showed that a monotone class of GAI rules can control online FDR as opposed to the modified FDR controlled in (Foster and Stine, 2008; Aharoni and Rosset, 2014). Within this class, of a special note is a procedure called *LORD* that performs consistently well in practice. Ramdas et al. (2017b) modified the GAI class (referred as to GAI++) to improve its statistical power (uniformly) while still controlling FDR, and the improved LORD++ method arguably represents the current state-of-the-art in the area. Very recently, (Ramdas et al., 2018) empirically demonstrated that using adaptiveness, some further improvements in the power over LORD++ can be obtained. In this paper, we mostly focus on GAI/GAI++ class, but certain results also carry over to the procedure of (Ramdas et al., 2018).

All above online testing procedures take p-values as input and make decisions based on previous 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 $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, which encodes contextual[1] information related to the tested hypothesis. The feature vector X_i only carries 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 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.

The problem of using side information in testing has been considered in offline setting (Ignatiadis et al., 2016; Genovese et al., 2006; Li and Barber, 2016; Ramdas et al., 2017a; Xia et al., 2017; Lei and Fithian, 2018). We review relevant prior work in offline setting in Appendix A. In this paper we focus on online setting, where 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 online testing problem has not been considered before. Our main contributions in this paper are as follows.

Incorporating Contextual Information. We propose a new broad class of powerful online testing rules, referred to as *contextual generalized alpha-investing* (CGAI) rules, which incorporates the available contextual features in the testing process. We also prove that any monotone rule from this class can control online FDR under some standard assumptions. Formally, we assume each hypothesis H is characterized by a tuple (P, X) where $P \in (0,1)$ is the p-value, and X is the contextual feature vector 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 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 time based on previous decisions and contextual information seen so far. The test for each hypothesis H_t takes the form $\mathbb{1}\{P_t \le \alpha_t\}$. Under the independence of p-values, and mutual independence between the p-values and the contextual features for null hypotheses, we show that any monotone rule from this class controls FDR below a preassigned level at any time. We also show that a variant of FDR (mFDR) can be controlled under a weaker assumption on p-values.

Context Weighting. We focus on a subclass of CGAI rules, referred to as *context-weighted generalized alpha-investing* (CwGAI) rules, for designing a practical online FDR control procedure. In particular, we take a parametric function $\omega(;\theta)$ with parameters θ, and at time t use $\omega(X_t;\theta)$ as a 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.

Statistical Power Analysis. We then look into the effect of context weighting in discovering true positives. Considering a general model of random weighting, and assuming that weights are positively associated with false null hypotheses, we derive a natural sufficient condition under which the weighting improves the power in an online setting, while still controlling FDR. We also discuss techniques for verifying this power improvement condition in practice. This is the first result that demonstrates the benefits of appropriate weighting in the online setting. Previously such results were only known in the offline setting (Genovese et al., 2006).

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

2 Related Online FDR Control Rules

We start with a review of online multiple testing model which was first introduced by (Foster and Stine, 2008). Considering a setting where an ordered (possibly infinite) sequence of hypotheses arriving in a stream, denoted by $\mathcal{H} = (H_1, H_2, H_3, \dots)$, we have to decide at each timestep t whether to reject H_t having only access to previous decisions. $H_t \in \{0, 1\}$ indicates if tth hypothesis is a true

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

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.,

$$\Pr[P_t \le u] \le u, \text{ for all } u \in [0,1]. \tag{1}$$

The marginal distribution of the p-values under alternative (non-null) hypotheses can be arbitrary. The only requirement is that they should be stochastically smaller than the uniform distribution, which means they carry signal that can differentiate them from nulls. Let $\mathcal{H}^0 = \{t : H_t = 0\}$ ($\mathcal{H}^1 = \{t : H_t = 1\}$) index the true (false) null hypotheses.

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

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

A rejection of the null hypothesis H_t indicated by the event $R_t = 1$ is also referred to as a *discovery*. Let us define 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 $\mathcal{H}(T) = (H_1, \dots, H_T)$. 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. Then the online false discovery proportion and rate till time T are defined as:

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

where $R(T) \vee 1 = \max\{R(T), 1\}$. The expectation is over the underlying randomness. Similarly, let $S(T) = \sum_{t \in \mathcal{H}^1} R_t$ be the number of true discoveries and let $N_1(T)$ be the number of true non-nulls till time T. Then online true discovery proportion and rate 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 to as *power*. In online hypothesis testing, our goal is to design a sequence of significance levels $(\alpha_t)_{t \in \mathbb{N}}$ such that we can control the online FDR at a desired 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 can be computed without the underlying true labels (ground truth). A variant of FDR studied in early online testing works (Foster and Stine, 2008) is the *marginal FDR*, defined as: $\text{mFDR}(T)_\eta = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)]+\eta}$, with a special case of $\text{mFDR}(T) = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)]+1}$ when $\eta = 1$. Note the gap between FDR and mFDR can be very significant, and controlling mFDR does not ensure controlling FDR at a similar level (Javanmard and Montanari, 2018). We also provide a guarantee on mFDR control in a contextual setting under weaker assumptions on p-values.

Generalized Alpha-Investing Rules. Foster and Stine (2008) proposed the first class of online multiple testing rules (referred to as alpha-investing rules) to control mFDR, which was extended by (Aharoni and Rosset, 2014) to generalized alpha-investing (GAI) rules. The GAI rules covers most of the online testing rules in the current literature.

Any rule of GAI class generates the significance level α_t at time t based on past decisions of the rule till time $t-1$: $\alpha_t = \alpha_t(R_1, \dots, R_{t-1})$. This means that α_t does not directly depend on the observed p-values but only on past decisions. Let $\mathcal{F}^t = \sigma(R_1, \dots, R_t)$ be the sigma-field of decisions till time t. In GAI rules, we require that $\alpha_t \in \mathcal{F}^{t-1}$.

Specifically, it begins with a wealth of $W(0) < \alpha$, and keeps track of the available wealth $W(t)$ after t steps. At each time t, an amount of ϕ_t, which is the *penalty* of testing the tth hypothesis at level α_t, will be deducted from the remaining wealth. If the tth hypothesis is rejected, i.e., $R_t = 1$, an extra wealth of amount ψ_t is *rewarded* to the current wealth. This can be explicitly stated as:

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

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

where w_0 and the nonnegative sequences $\alpha_t, \phi_t, \psi_t \in \mathcal{F}^{t-1}$ are user-defined. The wealth $W(t)$ is required to be always non-negative, and thus $\phi_t \le W(t-1)$. Once the wealth ever equals zero, the procedure is not allowed to make any further rejections since it has to set $\alpha_t = 0$ from then on. An additional restriction is needed for the goal to control FDR, in that the reward ψ_t has to be bounded whenever a rejection takes place. Formally, the constraints are:

$$\phi_t \le W(t-1), \tag{5}$$

$$\psi_t \le \min\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\}. \tag{6}$$

Javanmard and Montanari (2015, 2018) defined b_t as a user-chosen constant $b_0 = \alpha - w_0$ and proved the FDR control for monotone GAI rules under independence of p-values. The monotonicity of a rule is defined as:

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

$$\alpha_t(\tilde{R}_1, \dots, \tilde{R}_{t-1}) \le \alpha_t(R_1, \dots, R_{t-1}). \tag{7}$$

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, ρ_k defined as $\rho_k := \min_{i \in \mathbb{N}} \{\sum_{t=1}^{i} R_t = k\}$, is the time of kth rejection. Ramdas et al. (2017b) refer to this class of rules as GAI++ rules. Unless otherwise specified, we use this b_t (from GAI++) throughout this paper.

Level based On Recent Discovery (LORD) Rules. One popular subclass of GAI rules (proposed by (Javanmard and Montanari, 2015, 2018)) that is LORD, where significance level α_t is a function based only on *most recent discovery time*. Formally, we choose any sequence of non-increasing nonnegative constants $\gamma = (\gamma_t)_{t=1}^{\infty}$ 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., $\tau_t := \max\{i \in \{1, \dots, t-1\} : R_i = 1\}$, with $\tau_t = 0$ for all t before the first discovery. The LORD (Javanmard and Montanari, 2015, 2018) rule defines α_t, ϕ_t, ψ_t in the following generalized alpha-investing fashion.

$$\textbf{LORD: } W(0) = w_0,$$
$$\phi_t = \alpha_t = \begin{cases} \gamma_t w_0 & \text{if } t \le \rho_1 \\ \gamma_{t-\tau_t} b_0 & \text{if } t > \rho_1, \end{cases}$$
$$\psi_t = b_0 = \alpha - w_0.$$

Javanmard and 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), and we set $b_0 = w_0 = \alpha/2$, in which case, the above rule could be simplified as $\phi_t = \alpha_t = \gamma_{t-\tau_t} b_0$. As with any GAI rule, (Ramdas et al., 2017b) defined LORD++ by replacing b_0 with $b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t-1\}$ and showed it achieves a power increase while still controls online FDR at same level α. We describe LORD and LORD++ in little more detail in Appendix B. Note that both LORD and LORD++ rules satisfy the monotonicity condition from (7).

SAFFRON Procedure. This is a recent 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 is that SAFFRON is an adaptive method, based on adaptively estimating the proportion of true nulls, and can be viewed as an online extension of Storey's adaptive version of BH procedure from the offline setting. SAFFRON does not belong to the GAI class. See Appendix G for more details about SAFFRON, where we also extend the FDR control results of SAFFRON from (Ramdas et al., 2018) to a weighted version. Our experiments with SAFFRON suggests that contextual information could potentially help here too.

3 Contextual Online FDR Control

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 contextual features for each variant (e.g., its location, conservation, epigenetics, etc.) which could inform how likely the variant is to have a true association. Missing details from section are collected in Appendix C.

To deal with such situations, we now assume that a p-value $P_t \in (0,1)$ and a vector of contextual features $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$ are observed for each hypothesis H_t. At each step t, we have to decide whether to reject H_t having 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 correct discoveries by using the contextual information. Under the alternative, we denote the density (PDF) of p-values as $f_1(p \mid X)$ (depending on the feature vector $X \in \mathcal{X}$) and the corresponding cumulative distribution (CDF) 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 C.

Definition 1 (Contextual Online FDR Control)**.** *Given a (possibly infinite) sequence of (P_t, X_t)'s ($t \in \mathbb{N}$) where $P_t \in (0,1)$ and $X_t \in \mathcal{X}$, generate significance levels α_t's as a function of prior decisions and contextual features $\alpha_t = \alpha_t(R_1, \dots, R_{t-1}, X_1, \dots, X_t)$, and a corresponding set of decisions $R_t = \mathbb{1}\{P_t \le \alpha_t(R_1, \dots, R_{t-1}, X_1, \dots, X_t)\}$ such that $\sup_T FDR(T) \le \alpha$.*

We now define a contextual extension of GAI rules, that we refer to as *Contextual Generalized Alpha-Investing* (contextual GAI or CGAI) rules. In the presence of contextual information, we consider the sigma-field of decisions till time t as $\mathcal{F}^t = \sigma(R_1, \dots, R_t)$, and the sigma-field of features till time t as $\mathcal{G}^t = \sigma(X_1, \dots, X_t)$.

Definition 2 (Contextual GAI Rule)**.** *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, with the GAI conditions (3), (4), (5), (6) satisfied.*

We set $b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t-1\}$ as proposed by (Ramdas et al., 2017b). Similar to that in GAI rules (7), we define monotonicity property for contextual GAI rules as follows:

$$\textbf{Monotoncity: } \text{If } \tilde{R}_i \le R_i \text{ for all } i \le t-1, \text{ then}$$
$$\alpha_t(\tilde{R}_1, .., \tilde{R}_{t-1}, X_1, .., X_t) \le \alpha_t(R_1, .., R_{t-1}, X_1, .., X_t),$$
$$\text{for any fixed } \mathbf{X}^t = (X_1, .., X_t). \quad (8)$$

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

The following theorem establishes the FDR control for any monotone contextual GAI rule under an independence assumption between p-values and contextual features for the null hypotheses. As mentioned above, the p-values (P_t) could be arbitrary related to the contextual features (X_t) under the alternative (when $H_t = 1$). These assumptions are standard in multiple testing literature (see, e.g., (Ramdas et al., 2017b; Javanmard and Montanari, 2018; Xia

et al., 2017) among others).[2] 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 and 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 $\mathcal{F}^{t-1}$.

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 $(X_t)_{t\in\mathbb{N}}$ under the null (whenever $H_t = 0$), then for any monotone contextual generalized alpha-investing rule (satisfying conditions (3), (4), (5), (6), (8)), we have online FDR control,* $\sup_{T\in\mathbb{N}} \mathrm{FDR}(T) \le \alpha$.

Turning to mFDR, we can also prove a guarantee for mFDR control under a weaker condition than that in Theorem 1 by relaxing the independence assumptions to a weaker conditional super-uniformity assumption.

Conditional super-uniformity: If $H_t = 0$, then

$$\Pr[P_t \le \alpha_t \mid \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)] \le \alpha_t. \quad (9)$$

By definition of the marginal super-uniformity of p-values under the null (1), means that for independent p-values the conditional super-uniformity in (9) holds. So the assumption (9) is indeed weaker than the Theorem 1 assumptions. Our next theorem proves mFDR control for any contextual GAI rule (not necessarily monotone) under this weaker 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 (9)), then for any contextual generalized alpha-investing rule (satisfying conditions (3), (4), (5), (6)), we have online mFDR control,* $\sup_{T\in\mathbb{N}} \mathrm{mFDR}(T) \le \alpha$.

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

4 Context-weighted GAI Rules

The contextual GAI rules 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, it considers 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,

$$\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t) := \alpha_t(R_1, \ldots, R_{t-1}) \cdot \omega(X_t; \theta), \quad (10)$$

where $\omega(X_t; \theta)$ is a parametric *weight function* with parameters $\theta \in \Theta$. Since CwGAI is a subclass of CGAI rules, the above FDR and mFDR control theorems from previous section are valid for this class too. Applying this idea of context-weighting to LORD++ (resp. LORD) give rise to a new class of testing procedure that we refer to as CwLORD++ (resp. CwLORD) (defined in Appendix D).

Our reasons for considering this subclass include: (a) 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 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 context-weighted GAI rules when the weight function equals 1. We illustrate the relationship among various classes of testing rules in Figure 3 of Appendix D.

The idea of weighting p-values using prior information has been widely studied in offline multiple testing setup (Genovese et al., 2006; Ignatiadis et al., 2016; Li and Barber, 2016; Lei and Fithian, 2018; Xia et al., 2017; Ramdas et al., 2017a). 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. 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 true null which makes the procedure reject it more conservatively.

Weighting in Online vs. Offline Setting with FDR Control. In the offline setting, prior weights are usually rescaled to have unit mean, and then existing offline FDR control algorithm is applied to the weighted p-values P_i/ω_i instead of P_i (Genovese et al., 2006). However, in the online setting, the weights are computed at each timestep without knowing the total number of hypothesis or contextual information, thus cannot be rescaled to have unit mean in advance. Instead, as presented in (10), 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 subtle difference is that when α_t's are weighted, the

[2] Note that a standard assumption in hypothesis testing is that the p-values under the null are uniformly distributed in $(0, 1)$, which does not depend on the contextual features. That means the mutual independence of p-values and contextual features is valid.

penalty ϕ_t's and rewards ψ_t's are also adjusted according to the GAI constraints. For example, as dictated by (6), 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.

5 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 and 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 this section are collected in Appendix E.

We consider a mixture model where each null hypothesis is false with a fixed probability π_1, and the p-values are all independent. While the mixture model is *idealized*, it does offer a natural ground for comparing the power of various testing procedures (Genovese et al., 2006; Javanmard and Montanari, 2018). 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

$$\begin{gathered} H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1), \\ X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}), \\ P_t \mid H_t = 0, X_t \sim \text{Uniform}(0,1), \\ P_t \mid H_t = 1, X_t \sim F_1(p \mid X_t). \end{gathered}$$

where $0 < \pi_1 < 1$ and where $\mathcal{L}_0(\mathcal{X})$, $\mathcal{L}_1(\mathcal{X})$ are two probability distribution on the contextual feature space $\mathcal{X}$. Let $F = \int F_1(p \mid X) \mathrm{d}\mathcal{L}_1(\mathcal{X})$ be the marginal distribution of p-value under alternative. 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). We do not require that the contextual features X_t's be independent, but only that they be identically distributed as $\mathcal{L}_0(\mathcal{X})$ (under null) or $\mathcal{L}_1(\mathcal{X})$ (under alternative).

General Weighting Scheme. We consider the general weighting as in (Genovese et al., 2006) where weight is a random variable and conditionally independent of P_t given H_t. We assume that weight ω_t has different marginal distributions under null and alternative,

$$\omega_t \mid H_t = 0 \sim Q_0, \quad \omega_t \mid H_t = 1 \sim Q_1, \tag{11}$$

with Q_0, Q_1 unknown continuous distributions on $(0, \infty)$. Under the mixture setup,

$$\omega_t \overset{\text{i.i.d.}}{\sim} (1-\pi_1)Q_0 + \pi_1 Q_1, \tag{12}$$

with P_t and ω_t being conditionally independent given H_t for all $t = 1 \ldots, \infty$.

Contextual Weighting Scheme. This framework of weighting in (11) is very general. For example, it includes as a special case, the following contextual weighting scheme, where we assume that there exists a weight function of contextual features $\omega : \mathcal{X} \times \Theta \to \mathbb{R}$, and the distributions of weights are defined as:

$$\begin{aligned} \omega_t \mid H_t = 0 &\sim \omega(X;\theta), \text{ with } X \sim \mathcal{L}_0(\mathcal{X}), \\ \omega_t \mid H_t = 1 &\sim \omega(X;\theta), \text{ with } X \sim \mathcal{L}_1(\mathcal{X}). \end{aligned} \tag{13}$$

Now Q_0 and Q_1 are defined as the distributions of $\omega(X;\theta)$ under the null and alternative, respectively. Given Q_0 and Q_1, the weight ω_t is sampled as in (12).[3] Note that while the distributions Q_0 and Q_1 for weights are defined through X_t's distribution, the weight ω_t is sampled i.i.d. from the mixture model $(1-\pi_1)Q_0 + \pi_1 Q_1$, regardless of the value of X_t. Note that the independence assumption on p-values can still be satisfied even when the $X_t's$ are dependent.[4] Since this contextual weighting scheme is just a special case of the above general weighting scheme, in the remainder of this section, we work with the general weighting scheme.

Informativeness. Under (11), 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 assume that the weighting is *informative*, based on the following definition from (Genovese et al., 2006) in the offline setting,

$$u_0 < 1, \; u_1 > 1, \; u = \mathbb{E}[\omega] = (1-\pi_1)u_0 + \pi_1 u_1 = 1. \tag{14}$$

Remark 1. *Informative-weighting places a natural condition on the weights. Roughly it means that the weight should be positively associated to true alternatives (or the weight under alternative is more likely to be larger than that under the null). The marginal mean of weight $\mathbb{E}[\omega]$ is not necessary to be one. But for the theoretical power comparison of different procedures, it is convenient to scale the weight to have unit mean so that we can use the p-value reweighting akin to the offline setting. For empirical experiments, we will use an instantiation of CwLORD++ (see Section 6), that does not require the weight to have mean one.*

Comparison of Power. To compare different procedures, it is important to estimate their statistical power. Here we establish sufficient conditions under which a weighting could lead to a power increase for LORD. We work with (a version of) the popular LORD procedure from (Javanmard and Montanari, 2018), which sets

$$W(0) = w_0 = b_0 = \alpha/2, \; \phi_t = \alpha_t = b_0 \gamma_{t-\tau_t}, \; \psi_t = b_0. \tag{15}$$

[3] In case , $X_t \overset{\text{i.i.d.}}{\sim} (1-\pi_1)\mathcal{L}_0 + \pi_1 \mathcal{L}_1$, then one can define ω_t directly as $\omega_t = \omega(X_t;\theta)$ with Q_0 and Q_1 defined as the distributions of $\omega(X_t;\theta)$ under the null and alternative, respectively.

[4] In practice it is common that the contextual features are dependent (e.g., same genes or genetic variants may be tested in multiple independent experiments at different time), but as long as the tests are carried out independently the p-values are still independent.

As shown by (Javanmard and Montanari, 2018), the power of LORD, under the mixture model, almost surely equals[5]

$$\liminf_{T\to\infty} \text{TDP}(T) = \Big(\sum_{m=1}^{\infty}\prod_{j=1}^{m}(1-G(b_0\gamma_j))\Big)^{-1}, \qquad (16)$$

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* (15) *to the weighted p-values* $(P_1/\omega_1, P_2/\omega_2, \dots)$.

Weighted LORD is not strictly a contextual GAI rule (see discussion in Appendix E), however we establish FDR control of weighted LORD through Proposition 1(Appendix E). 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 $0 \le a < a_0$. The following theorem is based on comparing the power of weighted LORD (from Theorem 6 in Appendix E) with the power bound of LORD (16).

Theorem 3 (Power Separation)**.** *Suppose that the parameters in LORD* (15) *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 property in* (14)*. Then, the average power of weighted LORD is greater than or equal to that of LORD almost surely.*

The results show that using the informative context-weighting in the LORD rules will help in making more true discoveries. And we can also check the informative-weighting property by checking whether the mean of the weight distribution under alternative (Q_1) is greater than that of the corresponding distribution under null (Q_0), with details in Appendix F. Here the contextual features can be arbitrary with different dimensionality and our approach and results are still valid. We now conclude this section with a simple example of how the conditions of Theorem 3 are easily satisfied in a common statistical model.

Example 1. *To interpret the weight condition in Theorem 3, let's take a concrete example and consider the hypotheses* $(H_1, \dots, H_T)$ *concerning the means of normal distributions (referred to as* normal means model*) with test statistics* $Z_t \sim \mathcal{N}(\mu, 1)$*. So the two-sided p-values are* $P_t = 2\Phi(-|Z_t|)$*. Suppose under the null hypothesis* $\mu = 0$*, and under the alternative* $0 < \mu \le 4$*. Then we can compute that* $a_0 > 0.022$ *for any* μ *such that* $0 < \mu \le 4$*. In fact,* a_0 *increases as* μ *decreases. Setting* $\alpha = 0.05$*,* $T = 10^5$*, and* $\{\gamma_t\}_{t\in\mathbb{N}}$ *as suggested by (Javanmard and Montanari, 2018), we get that as long as the weight* ω *is less than* $a_0/(b_0\gamma_1) \approx 7.52$*, the condition for Theorem 3 is satisfied.*

6 Experimental Evaluation

Now we propose a practical procedure for contextual online FDR control based on context-weighted GAI rules, which sets $\alpha_t(R_1, \dots, R_{t-1}, X_1, \dots, X_t) := \alpha_t(R_1, \dots, R_{t-1}) \cdot \omega(X_t; \theta)$, and present numerical experiments to illustrate the performance with this procedure. In the following, we use $\alpha_t(X_t; \theta)$ as a short to represent $\alpha_t(R_1, \dots, R_{t-1}) \cdot \omega(X_t; \theta)$. Technically, we can use any parametric function $\omega(X_t; \theta)$ (with parameter set $\theta \in \Theta$) to model the weight function. Here we choose a deep neural network (multilayer perceptron) due to its expressive power, as in offline FDR control result by (Xia et al., 2017). It requires the input vectors to have the same length, but one can always replace it with other parametric models that can handle variable length inputs, and with that we get a scheme that can handle contextual features of different dimensionality. Given this, a natural goal is to find $\theta \in \Theta$ that maximizes the number of empirical discoveries (or discovery rate) while controlling the FDR. Note that if the function $\alpha_t(R_1, \dots, R_{t-1})$ is monotone (such as with LORD or LORD++) with respect to R_i's, the function $\alpha_t(R_1, \dots, R_{t-1}) \cdot \omega(X_t; \theta)$ is also monotone with respect to R_i's.

Training the Network, Setting θ. Given a stream $((P_t, X_t))_{t\in\mathbb{N}}$, the algorithm processes the stream in batches, in a single pass. Let $b \ge 1$ 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: $\text{EDR}_j = \sum_{i=jb+1}^{(j+1)b} \mathbb{1}\{P_i \le \alpha_i(X_i; \theta_j)\}/b$. Since the above function is not differentiable, we use the sigmoid function σ to approximate the indicator function, and define $\text{EDR}_j = \sum_{i=jb+1}^{(j+1)b} \sigma(\lambda(\alpha_i(X_i; \theta_j) - P_i))/b$. 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. We state the training procedure in Algorithm 1 (Appendix F).

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 by using LORD++ as the monotone GAI rule to set $\alpha_t(R_1, \dots, R_{t-1})$ in $\alpha_t(X_t; \theta)$.

Verifying Informativeness. We can add the verification of informative-weighting property (14) to the above procedure under various realistic scenarios such as in presence of feedback or a validation set. See discussion in Appendix F.

Experimental Results. We now discuss results for numerical experiments with both synthetic and real data to com-

[5] Javanmard and Montanari (2018) proposed multiple versions of LORD, and as noted by them, the bound in (16) lower bounds the power on all the versions of LORD under the mixture model.

pare the performance of our proposed CwLORD++ with a state-of-the-art online testing rule LORD++ (Ramdas et al., 2017b). We present the synthetic data experiments on the normal means model in Appendix F.1, with results showing that while FDR is controlled for both LORD++ and CwLORD++, the power of our CwLORD++ uniformly dominates that of LORD++. Our real data experiments focus on a diabetes prediction problem and gene expression data analyses. Experiments with the SAFFRON procedure are presented in Appendix G. The experimental code is also attached in supplementary material for reproducibility.

Diabetes Prediction Problem. We apply our online multiple testing rules to a real application of diabetes prediction, to test if patients are at risk of developing diabetes. For each patient i, the null hypothesis H_i is the "patient will not develop diabetes". Machine learning algorithms are now commonly used to construct predictive health scores for patients. A high predicted risk score can trigger an intervention (such as medical follow-up), which can be expensive and sometimes unnecessary, and thus it is important to control the fraction of false alerts. The dataset was released as part of a Kaggle competition[6], which contains de-identified medical records of 9948 patients. For each patient, we have a response variable 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.

We train a predictive score based on the available records, and then apply our online multiple testing rule rules to control FDR on test set. Our overall methodology is similar to that of (Javanmard and Montanari, 2018) in their FDR control experiments on this dataset. We regard the biographical information of patients as contextual features. Such choice is loosely based on the idea of *personalization* common in machine learning applications. Note that in theory, for our procedure, one could use any set of features as context. We describe the dataset and the ML training process in detail in Appendix F.2. Here are the results when $\alpha = 0.2$.

Table 1: Diabetes Dataset Results, $\alpha = 0.2$

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

Notice that FDR is under control for both procedures, and the power of CwLORD++ is substantially more (about 51%) than LORD++. This improvement shows the benefits of using contextual features for improving the power with FDR control in a machine learning setup. We probe the reasons for this improvement, and present results with different nominal FDR levels from 0.1 to 0.5 in Appendix F.2.

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.

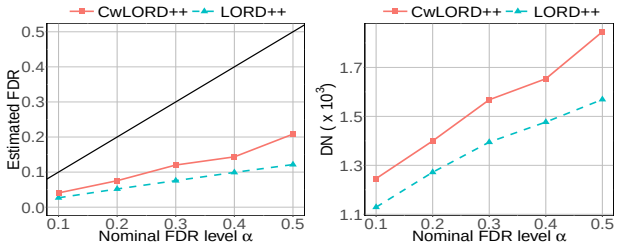

Figure 1: FDR and discovery numbers on Airway RNA-Seq.

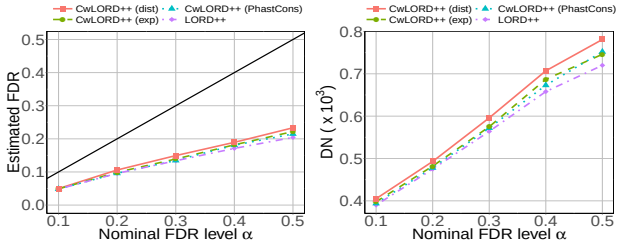

Figure 2: FDR and discovery numbers on GTEx.

The Airway RNA-Seq dataset contains $n = 33469$ genes, with the aim to identify glucocorticoid responsive (GC) genes that modulate cytokine function in airway smooth muscle cells. The p-values are obtained in two-sample differential analysis of gene expression levels. Log counts of each gene serves as the contextual feature. Figure 1 reports the empirical FDR and the discovery number. We see that our CwLORD++ procedure make about 10% more discoveries than the LORD++ procedure.

The GTEx study is to quantify the expression Quantitative Trait Loci (eQTLs) in human tissues, where 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. We consider 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 2 reports the empirical FDR and the discovery number where for CwLORD++ with different contextual features. Results show that CwLORD++ has an increase in discovery number by 5.5%, 2.6%, 2.9% using GTEx-dist, GTEx-exp, and GTEx-PhastCons as the contextual feature respectively, compared to the LORD++ procedure. We also provide experimental results with multi-dimensional feature vectors in Appendix F.3, and draw a similar conclusion.

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

[7]https://www.dropbox.com/sh/wtp58wd60980d6b/AAA4wA60ykP-fDfS5BNsNkiGa?dl=0.

References

Aharoni, E. and 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.

Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001.

Benjamini, Y. and 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. and 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., and Cox, D. R. (1967). *Renewal theory*, volume 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., and Owen, A. B. (2015). Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika*, 102(4):753–766.

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

Foster, D. P. and 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. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

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

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

G'Sell, M. G., Wager, S., Chouldechova, A., and 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., and 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., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577.

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

Javanmard, A. and 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. and Fithian, W. (2016). Power of ordered hypothesis testing. *arXiv preprint arXiv:1606.01969*.

Lei, L. and 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. and Barber, R. F. (2016). Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*.

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

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

Ramdas, A., Zrnic, T., Wainwright, M., and 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., and Tse, D. (2017). Neuralfdr: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, pages 1541–1550.