The Annals of Statistics
2018, Vol. 46, No. 2, 526–554
https://doi.org/10.1214/17-AOS1559

ONLINE RULES FOR CONTROL OF FALSE DISCOVERY RATE AND FALSE DISCOVERY EXCEEDANCE

BY ADEL JAVANMARD[1] AND ANDREA MONTANARI[2]

University of Southern California and Stanford University

Multiple hypothesis testing is a core problem in statistical inference and arises in almost every scientific field. Given a set of null hypotheses $\mathsf{H}(n) = (H_1, \dots, H_n)$, Benjamini and Hochberg [*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** (1995) 289–300] introduced the false discovery rate (FDR), which is the expected proportion of false positives among rejected null hypotheses, and proposed a testing procedure that controls FDR below a preassigned significance level. Nowadays FDR is the criterion of choice for large-scale multiple hypothesis testing.

In this paper we consider the problem of controlling FDR in an *online manner*. Concretely, we consider an ordered—possibly infinite—sequence of null hypotheses $\mathsf{H} = (H_1, H_2, H_3, \dots)$ where, at each step i, the statistician must decide whether to reject hypothesis H_i having access only to the previous decisions. This model was introduced by Foster and Stine [*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** (2008) 429–444].

We study a class of *generalized alpha investing* procedures, first introduced by Aharoni and Rosset [*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** (2014) 771–794]. We prove that any rule in this class controls online FDR, provided p-values corresponding to true nulls are independent from the other p-values. Earlier work only established mFDR control. Next, we obtain conditions under which generalized alpha investing controls FDR in the presence of general p-values dependencies. We also develop a modified set of procedures that allow to control the false discovery exceedance (the tail of the proportion of false discoveries). Finally, we evaluate the performance of online procedures on both synthetic and real data, comparing them with offline approaches, such as adaptive Benjamini–Hochberg.

1. Introduction.

The common practice in claiming a scientific discovery is to support such claim with a p-value as a measure of statistical significance. Hypotheses with p-values below a significance level α, typically 0.05, are considered to be *statistically significant*. While this ritual controls type I errors for single testing problems, in case of testing multiple hypotheses it leads to a large number

Received March 2016; revised January 2017.

[1]Supported in part by a CSoI fellowship during the course of this work (NSF Grant CCF-0939370).

[2]Supported in part by NSF Grants CCF-1319979 and DMS-1106627, and the AFOSR Grant FA9550-13-1-0036.

MSC2010 subject classifications. Primary 62F03, 62F05; secondary 62L99.

Key words and phrases. Hypothesis testing, false discovery rate (FDR), false discovery exceedance (FDX), online decision making.

of false positives (false discoveries). Consider, for instance, a setting in which N hypotheses are to be tested, but only a few of them, say s, are non-null. If we test all of the hypotheses at a fixed significance level α, each of $N - s$ truly null hypotheses can be falsely rejected with probability α. Therefore, the number of false discoveries—equal to $\alpha(N - s)$ in expectation—can substantially exceed the number s of true non-nulls.

The false discovery rate (FDR)—namely, the expected fraction of discoveries that are false positives—is the criterion of choice for statistical inference in large scale hypothesis testing problem. In their groundbreaking work [4], Benjamini and Hochberg (BH) developed a procedure to control FDR below a pre-assigned level, while allowing for a large number of true discoveries when many non-nulls are present. The BH procedure remains—with some improvements—the state-of-the-art in the context of multiple hypothesis testing, and has been implemented across genomics [34], brain imaging [12], marketing [32], and many other applied domains.

Standard FDR control techniques, such as the BH procedure [4], require aggregating p-values for all the tests and processing them jointly. This is impossible in a number of applications which are best modeled as an online hypothesis testing problem [11] (a more formal definition will be provided below):

Hypotheses arrive sequentially in a stream. At each step, the analyst must decide whether to reject the current null hypothesis without having access to the number of hypotheses (potentially infinite) or the future p-values, but solely based on the previous decisions.

This is the case, for instance, with publicly available datasets, where new hypotheses are tested in an on-going fashion by different researchers [1]. Similar constraints arise in marketing research, where multiple A-B tests are carried out on an ongoing fashion [32]. Finally, scientific research as a whole suffers from the same problem: a stream of hypotheses are tested on an ongoing basis using a fixed significance level, thus leading to large numbers of false positives [17]. We refer to Section 1.2 for further discussion.

In order to illustrate the online scenario, consider an approach that would control the family-wise error rate (FWER), that is, the probability of rejecting at least one true null hypothesis. Formally,

$$\text{FWER}(n) \equiv \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}_{\theta}\big(V^{\theta}(n) \geq 1\big), \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters (including the set of non-null hypotheses) and $V^{\theta}(n)$ the number of false positives among the first n hypotheses. This metric can be controlled by choosing different significance levels α_i for tests H_i, with $\boldsymbol{\alpha} = (\alpha_i)_{i \geq 1}$ summable, for example, $\alpha_i = \alpha 2^{-i}$. Notice that the analyst only needs to know the number of tests performed before the current one, in order to implement this scheme. However, this method leads to small statistical power. In particular, making a discovery at later steps becomes very unlikely.

In contrast, the BH procedure assumes that all the p-values are given a priori. Given p-values $p_1, p_2, \dots, p_N$ and a significance level α, BH follows the steps below:

1. Let $p_{(i)}$ be the ith p-value in the (increasing) sorted order, and define $p_{(0)} = 0$. Further let

$$i_{\mathrm{BH}} \equiv \max\{0 \le i \le N : p_{(i)} \le \alpha i/N\}. \tag{2}$$

2. Reject H_j for every test with $p_j \le p_{(i_{\mathrm{BH}})}$.

As mentioned above, BH controls the false discovery rate defined as

$$\mathrm{FDR}(N) \equiv \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\theta}\left(\frac{V^{\theta}(N)}{R(N) \vee 1}\right), \tag{3}$$

where $R(N)$ is the total the number of rejected hypotheses. Note that BH requires the knowledge of *all* p-values to determine the significance level for testing the hypotheses. Hence, it does not address the online scenario.

In this paper, we study methods for *online* control of false discovery rate. Namely, we consider a sequence of hypotheses $H_1, H_2, H_3, \dots$ that arrive sequentially in a stream, with corresponding p-values $p_1, p_2, \dots$. We aim at developing a testing mechanism that ensures false discovery rate remains below a pre-assigned level α. A testing procedure provides a sequence of significance levels α_i, with decision rule

$$R_i = \begin{cases} 1, & \text{if } p_i \le \alpha_i \text{ (reject } H_i), \\ 0, & \text{otherwise (accept } H_i). \end{cases} \tag{4}$$

In *online* testing, we require significance levels to be functions of prior outcomes:

$$\alpha_i = \alpha_i(R_1, R_2, \dots, R_{i-1}). \tag{5}$$

Foster and Stine [11] introduced the above setting and proposed a class of procedures named *alpha investing rules*. Alpha investing starts with an initial wealth, at most α, of allowable false discovery rate. The wealth is spent for testing different hypotheses. Each time a discovery occurs, the alpha investing procedure earns a contribution toward its wealth to use for further tests. Foster and Stine [11] proved that alpha investing rules control a modified metric known as mFDR, defined as

$$\mathrm{mFDR}_{\eta}(n) \equiv \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\mathbb{E}(V^{\theta}(n))}{\mathbb{E}(R(n)) + \eta}. \tag{6}$$

In words, mFDR is the ratio of the expected number of false discoveries to the expected number of discoveries. As illustrated in the Supplementary Material [22], mFDR and FDR can be very different in situations with high variability. While FDR is the expected proportion of false discoveries, mFDR is the ratio of two expectations and hence is not directly related to any single sequence quantity.

Several recent papers [15, 27, 29] consider a "sequential hypothesis testing" problem that arises in connection with sparse linear regression. Let us emphasize that the problem treated in [15, 29] is substantially different from the one analyzed here. For instance, as discussed in Section 1.2, the methods of [15] achieve vanishingly small statistical power for the present problem.

1.1. *Contributions*. In this paper, we study a class of procedures that are known as *generalized alpha investing*, and were first introduced by Aharoni and Rosset in [1]. As in alpha investing [11], generalized alpha investing makes use of a potential sequence (wealth) that increases every time a null hypothesis is rejected, and decreases otherwise. However: (i) The pay-off and pay-out functions are general functions of past history; (ii) The pay-out is not tightly determined by the testing level α_i. This additional freedom allows to construct interesting new rules.

The contributions of this paper are summarized as follows.

Online control of FDR. We prove that generalized alpha investing rules control FDR, under the assumption of independent p-values, and provided they are *monotone* (a technical condition defined in the sequel). To the best of our knowledge, this is the first work[3] that guarantees online control of FDR.

Online control of FDR for dependent p-values. Dependencies among p-values can arise for multiple reasons. For instance, the same data can be re-used to test a new hypothesis, or the choice of a new hypothesis can depend on the past outcomes. We present a general upper bound on the FDR for dependent p-values under generalized alpha investing.

False discovery exceedance. FDR can be viewed as the expectation of false discovery proportion (FDP). In some cases, FDP may not be well represented by its expectation, for example, when the number of discoveries is small. In these cases, FDP might be sizably larger than its expectation with significant probability. In order to provide tighter control, we develop bounds on the false discovery exceedance (FDX), that is, on the tail probability of FDP.

Statistical power. In order to compare different procedures, we develop lower bounds on fraction of non-null hypotheses that are discovered (statistical power), under a mixture model where each null hypothesis is false with probability π_1, for a fixed arbitrary π_1.

We focus in particular on a concrete example of generalized alpha investing rule (called LORD below) that we consider particularly compelling. We use our lower bound to guide the choice of parameters for this rule.

Numerical validation. We validate our procedures on synthetic and real data in Section 5 and the Supplementary Material [22], showing that they control FDR and mFDR in an online setting. We further compare them with BH and Bonferroni

[3]Special cases were presented in our earlier technical report [21].

procedures. We observe that generalized alpha investing procedures can benefit from ordering of hypotheses. Specifically, they can achieve higher statistical power compared to offline benchmarks such as adaptive BH, when fraction of non-nulls is small and hypotheses can be a priori ordered in such a way that those most *likely* to be rejected appear first in the sequence.

1.2. *Further related work. General context.* An increasing effort was devoted to reducing the risk of fallacious research findings. Some of the prevalent issues such as publication bias, lack of replicability and multiple comparisons on a dataset were discussed in Ioannidis's 2005 papers [16, 17] and in [33].

Statistical databases. Concerned with the above issues and the importance of data sharing in the genetics community, [35] proposed an approach to public database management, called Quality Preserving Database (QPD). A QPD makes a shared data resource amenable to perpetual use for hypothesis testing while controlling FWER and maintaining statistical power of the tests. In this scheme, for testing a new hypothesis, the investigator should pay a price in form of additional samples that should be added to the database. The number of required samples for each test depends on the required effect size and the power for the corresponding test. A key feature of QPD is that type I errors are controlled at the management layer and the investigator is not concerned with p-values for the tests. Instead, investigators provide effect size, assumptions on the distribution of the data, and the desired statistical power. A critical limitation of QPD is that all samples, including those currently in the database and those that will be added, are assumed to have the same quality and are coming from a common underlying distribution. Motivated by similar concerns in practical data analysis, [9] applies insights from differential privacy to efficiently use samples to answer adaptively chosen estimation queries. These papers however do not address the problem of controlling FDR in online multiple testing.

Online feature selection. Building upon alpha investing procedures, [28] develops VIF, a method for feature selection in large regression problems. VIF is accurate and computationally very efficient; it uses a one-pass search over the pool of features and applies alpha investing to test each feature for adding to the model. VIF regression avoids overfitting due to the property that alpha investing controls mFDR. Similarly, one can incorporate LORD in VIF regression to perform fast online feature selection and provably avoid overfitting.

High-dimensional and sparse regression. There has been significant interest over the last two years in developing hypothesis testing procedures for high-dimensional regression, especially in conjunction with sparsity-seeking methods. Procedures for computing p-values of low-dimensional coordinates were developed in [18–20, 39, 41]. Sequential and selective inference methods were proposed in [10, 29, 38]. Methods to control FDR were put forward in [2, 7].

As exemplified by VIF regression, online hypothesis testing methods can be useful in this context as they allow to select a subset of regressors through a one-pass procedure. Also they can be used in conjunction with the methods of [29],

where a sequence of hypothesis is generated by including an increasing number of regressors (e.g., sweeping values of the regularization parameter).

In particular, [15, 27] develop multiple hypothesis testing procedures for ordered tests. Note, however, that these approaches fall short of addressing the issues we consider, for several reasons: (i) They are not online, since they reject the first $\hat{k}$ null hypotheses, where $\hat{k}$ depends on all the p-values. (ii) They require knowledge of all past p-values (not only discovery events) to compute the current score. (iii) Since they are constrained to reject all hypotheses before $\hat{k}$, and accept them after, they cannot achieve any discovery rate increasing with n, let alone nearly linear in n. For instance, in the mixture model of Section 4, if the fraction of true non-null is $\pi_1 < \alpha$, then the methods of [15, 27] achieves $O(1)$ discoveries out of $\Theta(n)$ true non-null. In other words, their power is of order $1/n$ in this simple case.

1.3. *Notations.* Throughout the paper, we typically use upper case symbols (e.g., $X, Y, Z, \ldots$) to denote random variables, and lower case symbols for deterministic values (e.g., $x, y, z, \ldots$). Vectors are denoted by boldface, for example, $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$ for random vectors, and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$ for deterministic vectors. Given a vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, we use $\boldsymbol{X}_i^j = (X_i, X_{i+1}, \ldots, X_j)$ to denote the subvector with indices between i and j. We will often consider sequences indexed by the same "time index" as for the hypotheses $\{H_1, H_2, H_3, \ldots\}$. Given such a sequence $(X_i)_{i\in\mathbb{N}}$, we denote by $X(n) \equiv \sum_{i=1}^n X_i$ its partial sums.

We denote the standard Gaussian density by $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, and the Gaussian distribution function by $\Phi(x) = \int_{-\infty}^x \phi(t)\,\mathrm{d}t$. We use the standard big-O notation. In particular, $f(n) = O(g(n))$ as $n \to \infty$ if there exists a constant $C > 0$ such that $|f(n)| \le Cg(n)$ for all n large enough. We also use $\sim$ to denote asymptotic equality, that is, $f(n) \sim g(n)$ as $n \to \infty$, means $\lim_{n\to\infty} f(n)/g(n) = 1$. We further use $\asymp$ for equality up to constants, that is, if $f(n) = \Theta(g(n))$, then there exist constants $C_1, C_2 > 0$ such that $C_1|g(n)| \le f(n) \le C_2|g(n)|$ for all n large enough.

2. Generalized alpha investing. In this section, we define generalized alpha investing rules, and provide some concrete examples. Our definitions and notations follow the paper of Aharoni and Rosset that first introduced generalized alpha investing [1].

2.1. *Definitions.* Given a sequence of input p-values $(p_1, p_2, \ldots)$, a *generalized alpha investing* rule generates a sequence of decisions $(R_1, R_2, \ldots)$ (here $R_j \in \{0, 1\}$ and $R_j = 1$ is to be interpreted as rejection of null hypothesis H_j) by using test levels $(\alpha_1, \alpha_2, \alpha_3, \ldots)$. After each decision j, the rule updates a potential function $W(j)$ as follows:

- If hypothesis j is accepted, then the potential function is decreased by a pay-out φ_j.

- If hypothesis j is rejected, then the potential is increased by an amount $\psi_j - \varphi_j$.

In other words, the pay-out φ_j is the amount paid for testing a new hypothesis, and the pay-off ψ_j is the amount earned if a discovery is made at that step.

Formally, a generalized alpha investing rule is specified by three (sequences of) functions $\alpha_j, \varphi_j, \psi_j : \{0,1\}^{j-1} \to \mathbb{R}_{\geq 0}$, determining test levels, pay-out and pay-off. Decisions are taken by testing at level α_j

$$R_j = \begin{cases} 1, & \text{if } p_j \leq \alpha_j = \alpha_j(R_1, \dots, R_{j-1}), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The potential function is updated via

$$W(0) = w_0, \tag{8}$$

$$W(j) = W(j-1) - \varphi_j(\boldsymbol{R}_1^{j-1}) + R_j \psi_j(\boldsymbol{R}_1^{j-1}), \tag{9}$$

with $w_0 \geq 0$ an initial condition. Notice in particular that $W(j)$ is a function of $(R_1, \dots, R_j)$.

A valid generalized alpha investing rule is required to satisfy the following conditions, for a constant $b_0 > 0$:

G1. For all $j \in \mathbb{N}$ and all $\boldsymbol{R}_1^{j-1} \in \{0,1\}^{j-1}$, letting $\psi_j = \psi_j(\boldsymbol{R}_1^{j-1})$, $\varphi_j = \varphi_j(\boldsymbol{R}_1^{j-1})$, $\alpha_j = \alpha_j(\boldsymbol{R}_1^{j-1})$, we have

$$\psi_j \leq \varphi_j + b_0, \tag{10}$$

$$\psi_j \leq \frac{\varphi_j}{\alpha_j} + b_0 - 1, \tag{11}$$

$$\varphi_j \leq W(j-1). \tag{12}$$

G2. For all $j \in \mathbb{N}$, and all $\boldsymbol{R}_1^{j-1} \in \{0,1\}^{j-1}$, if $W(j-1) = 0$ then $\alpha_j = 0$.

Notice that Condition (12) and G2 are well posed since $W(j-1)$, φ_j and α_j are functions of $\boldsymbol{R}_1^{j-1}$. Further, because of (12), the function $W(j)$ remains nonnegative for all $j \in \mathbb{N}$.

We later show that generalized alpha investing guarantees FDR control as a function of b_0 and w_0.

Throughout, we shall denote by $\mathcal{F}_j$ the σ-algebra generated by the random variables $\{R_1, \dots, R_j\}$.

DEFINITION 2.1. For $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n$, we write $\boldsymbol{x} \preceq \boldsymbol{y}$ if $x_j \leq y_j$ for all $j \in \{1, \dots, n\}$. We say that an online rule is *monotone* if the functions α_j are monotone nondecreasing with respect to this partial ordering [i.e., if $\boldsymbol{x} \preceq \boldsymbol{y}$ implies $\alpha_j(\boldsymbol{x}) \leq \alpha_j(\boldsymbol{y})$].

REMARK 2.2. Our notation differs from [1] in one point, namely we use w_0 for the initial potential (which is denoted by $\alpha\eta$ in [1]) and b_0 for the constant appearing in equations (10), (11) (which is denoted by α in [1]). We prefer to reserve α for the FDR level.[4]

REMARK 2.3. In a generalized alpha investing rule, as we reject more hypotheses the potential $W(j)$ increases and hence we can use large test levels α_j. In other words, the burden of proof decreases as we reject more hypotheses. This is similar to the BH rule, where the most significant p-values is compared to a Bonferroni cutoff, the second most significant to twice this cutoff and so on.

2.2. *Examples*. Generalized α-investing rules comprise a large variety of online hypothesis testing methods. We next describe some specific subclasses that are useful for designing specific procedures.

2.2.1. *Alpha investing*. Alpha investing, introduced by Foster and Stine [11], is a special case of generalized alpha investing rule. In this case, the potential is decreased by $\alpha_j/(1-\alpha_j)$ if hypothesis H_j is not rejected, and increased by a fixed amount b_0 if it is rejected. In formula, the potential evolves according to

$$W(j) = W(j-1) - (1-R_j)\frac{\alpha_j}{1-\alpha_j} + R_j b_0. \tag{13}$$

This fits the above framework by defining $\varphi_j = \alpha_j/(1-\alpha_j)$ and $\psi_j = b_0 + \alpha_j/(1-\alpha_j)$. Note that this rule depends on the choice of the test levels α_j, and of the parameter b_0. The test levels α_j can be chosen arbitrarily, provided that they satisfy condition (12), which is equivalent to $\alpha_j/(1-\alpha_j) \le W(j-1)$.

2.2.2. *Alpha spending with rewards*. Alpha spending with rewards was introduced in [1], as a special subclass of generalized alpha investing rules, which are convenient for some specific applications.

In this case, test levels are chosen to be proportional to the pay-out function, $\alpha_j = \varphi_j/\kappa$, with a proportionality coefficient κ. Conditions (10) and (11) coincide with[5]

$$0 \le \psi_j \le \min(\kappa\alpha_j + b_0, \kappa - 1 + b_0). \tag{14}$$

The choice of penalties φ_j is arbitrary as long as constraint (12) is satisfied. For instance, [1] uses $\varphi_j = c_1 W(j-1)$ with $c_1 \in (0,1)$.

[4]The use of η in [1] was related to control of mFDR_η in that paper.

[5]Note that [1] rescales the potential function by κ, and hence the condition on ψ_j is also rescaled.

2.2.3. *LORD.* As a running example, we shall use a simple procedure that we term LORD, for Levels based On Recent Discovery. LORD is easily seen to be a special case of alpha spending with rewards, for $\kappa = 1$.

Below, we present three different versions of LORD . For a concrete exposition, choose any sequence of nonnegative numbers $\boldsymbol{\gamma} = (\gamma_i)_{i\in\mathbb{N}}$, which is monotone nonincreasing (i.e., for $i \le j$ we have $\gamma_i \ge \gamma_j$) and such that $\sum_{i=1}^{\infty} \gamma_i = 1$. We refer to Section 4 for concrete choices of this sequence.

At each time i, we let $T(i)$ be the set of discovery times up to time i. We further define τ_i as the last time a discovery was made before i:

$$T(i) = \{\ell \in \{1, \dots, i-1\} : R_\ell = 1\} \qquad \tau_i = \max\{\ell : \ell \in T(i)\}.$$

At each step, if a discovery is made, we add an amount b_0 to the current wealth. Otherwise, we remove an amount of the current test level from the wealth. Formally, we set

$$W(0) = w_0, \qquad \psi_i = b_0, \qquad \varphi_i = \alpha_i, \tag{15}$$

where $\{W(j)\}_{j\ge 0}$ is defined recursively via equation (9).

We present three versions of LORD which differ in the way that the test levels α_i are set.

- LORD 1: We set the test levels solely based on the time of the last discovery. Specifically,

$$\alpha_i = \begin{cases} \gamma_i w_0 & \text{if } i \le t_1, \\ \gamma_{i-\tau_i} b_0 & \text{if } i > t_1, \end{cases} \tag{16}$$

  where t_1 denotes the time of first discovery. In words, up until the first discovery is made, we set levels by discounting the initial wealth, that is, $\gamma_i w_0$. After the first discovery is made, we use a fraction $\gamma_{i-\tau_i}$ of b_0 to spend in testing null hypothesis H_i.

- LORD 2: We set the test levels based on the previous discovery times. Specifically,

$$\alpha_i = \gamma_i w_0 + \Big(\sum_{\ell\in T(i)} \gamma_{i-\ell}\Big) b_0. \tag{17}$$

- LORD 3: In this alternative, the significance levels α_i depend on the past only through the time of the last discovery, and the wealth accumulated at that time. Specifically,

$$\alpha_i = \gamma_{i-\tau_i}\, W(\tau_i)\,, \tag{18}$$

Note that τ_i and $T(i)$ are measurable on $\mathcal{F}_{i-1}$, and hence in all the variants of LORD , φ_i, ψ_i are functions of $\boldsymbol{R}_1^{i-1}$, while $W(i)$ is a function of $\boldsymbol{R}_1^{i}$. Therefore, each of the above rules defines an online multiple hypothesis testing procedure. Indeed, the three versions of LORD are generalized alpha investing rules. Further, LORD 1 and LORD 2 are monotone rules (see Definition 2.1), while LORD 3 is not necessarily a monotone rule without making further assumptions on $\boldsymbol{\gamma}$.

3. Control of false discovery rate.

3.1. FDR *control for independent test statistics*. As already mentioned, we are interested in testing a—possibly infinite—sequence of null hypotheses $\mathcal{H} = (H_i)_{i\in\mathbb{N}}$. The set of first n hypotheses will be denoted by $\mathcal{H}(n) = (H_i)_{1\le i\le n}$. Without loss of generality, we assume H_i concerns the value of a parameter θ_i, with $H_i = \{\theta_i = 0\}$. Rejecting the null hypothesis H_i can be interpreted as θ_i being significantly nonzero. We will denote by Θ the set of possible values for the parameters θ_i, and by $\boldsymbol{\Theta} = \Theta^{\mathbb{N}}$ the space of possible values of the sequence $\boldsymbol{\theta} = (\theta_i)_{i\in\mathbb{N}}$.

Under the null hypothesis $H_i : \theta_i = 0$, the corresponding p-value is uniformly random in $[0, 1]$:

$$p_i \sim \mathsf{Unif}\big([0,1]\big). \tag{19}$$

Recall that R_i is the indicator that a discovery is made at time i, and $R(n) = \sum_{i=1}^n R_i$ the total number of discoveries up to time n. Analogously, let $V_i^{\boldsymbol{\theta}}$ be the indicator that a false discovery occurs at time i and $V^\theta(n) = \sum_{i=1}^n V_i^\theta$ the total number of false discovery up to time n. Throughout the paper, superscript θ is used to distinguish unobservable variables such as $V^\theta(n)$, from statistics such as $R(n)$. However, we drop the superscript when it is clear from the context.

There are various criteria of interest for multiple testing methods. We will mostly focus on the *false discovery rate (FDR)* [4], and we repeat its definition here for the reader's convenience. We first define the *false discovery proportion* (FDP) as follows. For $n \ge 1$,

$$\mathrm{FDP}^\theta(n) \equiv \frac{V^\theta(n)}{R(n)\vee 1}. \tag{20}$$

The false discovery rate is defined as

$$\mathrm{FDR}(n) \equiv \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_\theta\big(\mathrm{FDP}^\theta(n)\big). \tag{21}$$

Our first result establishes FDR control for all monotone generalized alpha investing procedures, where the monotonicity of a testing rule is given by Definition 2.1. Its proof is presented in the Supplementary Material [22].

THEOREM 3.1. *Assume the p-values $(p_i)_{i\in\mathbb{N}}$ to be independent. Then, for any monotone generalized alpha investing rule with $w_0 + b_0 \le \alpha$, we have*

$$\sup_n \mathrm{FDR}(n) \le \alpha. \tag{22}$$

The same holds if only the p-values corresponding to true nulls are mutually independent, and independent from the nonnull p-values.

By applying Theorem 3.1, we obtain that LORD 1 and LORD 2 control FDR at level α, as long as $w_0 + b_0 \le \alpha$. For LORD 3, such result cannot be obtained directly from Theorem 3.1 because this rule is not necessarily a monotone rule without making further assumptions on the sequence $\boldsymbol{\gamma}$. Nevertheless, in our numerical experiments, we focus on LORD 3 and as we show empirically, it also controls FDR.

REMARK 3.2. In the Supplementary Material [22], we prove a somewhat stronger version of Theorem 3.1, namely $\mathrm{FDR}(n) \le b_0 \mathbb{E}\{R(n)/(R(n) \vee 1)\} + w_0 \mathbb{E}\{1/(R(n) \vee 1)\}$. In particular, $\mathrm{FDR}(n) \,\square\, b_0$ when the total number of discoveries $R(n)$ is large, with high probability. This is the case for instance, when the hypotheses to be tested comprise a large number of "strong signals" (even if these form a small proportion of the total number of hypotheses).

Another possible strengthening of Theorem 3.1 is obtained by considering a new metric, that we call $\mathrm{sFDR}_\eta(n)$ (for smoothed FDR):[6]

$$\mathrm{sFDR}_\eta(n) \equiv \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left\{ \frac{V^{\boldsymbol{\theta}}(n)}{R(n) + \eta} \right\}. \tag{23}$$

The following theorem bounds $\mathrm{sFDR}_{w_0/b_0}(n)$ for monotone generalized alpha investing rules (cf. Definition 2.1).

THEOREM 3.3. *Under the assumptions of Theorem* 3.1, *for any* $w_0, b_0 > 0$, *we have*

$$\sup_n \mathrm{sFDR}_{w_0/b_0}(n) \le b_0. \tag{24}$$

Note that equation (24) implies (22) by using $R(n) + (w_0/b_0) \le (b_0 + w_0)R(n)/b_0$ for $R(n) \ge 1$. Also, $\mathbb{E}\{V^{\boldsymbol{\theta}}(n)/(R(n) + (w_0/b_0))\} \approx \mathrm{FDR}(n)$ if $R(n)$ is large with high probability.

Let us emphasize that the guarantee in Theorem 3.3 is different from the one in [1, 11], which instead use $\mathrm{mFDR}_\eta(n) \equiv \mathbb{E}\{V^{\boldsymbol{\theta}}(n)\}/(\mathbb{E}\{R(n)\} + \eta)$. As mentioned earlier, mFDR does not correspond to a single-sequence property.

REMARK 3.4. In the Supplementary Material [22], we show that Theorems 3.1 and 3.3 cannot be substantially improved, unless specific restrictions are imposed on the generalized alpha investing rule. In particular, we prove that there exist generalized alpha investing rules for which $\liminf_{n\to\infty} \mathrm{FDR}(n) \ge b_0$, and $\lim_{n\to\infty} \mathrm{sFDR}_{w_0/b_0} = b_0$.

We conclude that LORD is a monotone generalized alpha investing rule, as per Definition 2.1.

[6]Some authors [3] refer to this quantity as "modified FDR." We will not follow this terminology since its acronym (mFDR) gets confused with "marginal FDR" [1, 11].

3.2. FDR *control for dependent test statistics*. In some applications, the assumption of independent p-values is not warranted. This is the case—for instance—of multiple related hypotheses being tested on the same experimental data. Benjamini and Yekutieli [5] introduced a property called *positive regression dependency from a subset* I_0 (PRDS on I_0) to capture a positive dependency structure among the test statistics. They showed that if the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, then BH controls FDR. (See Theorem 1.3 in [5].) Further, they proved that BH controls FDR under general dependency if its threshold is adjusted by replacing α with $\alpha/(\sum_{i=1}^{N} \frac{1}{i})$ in equation (2).

Our next result establishes an upper bound on the FDR of generalized alpha investing rules, under general p-values dependencies. For a given generalized alpha investing rule, let $\mathcal{R}_i \equiv \{\boldsymbol{r}_1^i \in \{0,1\}^i : \mathbb{P}(\boldsymbol{R}_1^i = \boldsymbol{r}_1^i) > 0\}$, the set of decision sequences that have nonzero probability.

DEFINITION 3.5. An *index sequence* is a sequence of deterministic functions $\mathcal{I} = (\mathcal{I}_i)_{i\in\mathbb{N}}$ with $\mathcal{I}_i : \{0,1\}^i \to \mathbb{R}_{\geq 0}$. For an index sequence $\mathcal{I}$, let

$$R_i^{\mathrm{L}}(s) \equiv \min_{\boldsymbol{r}_1^{i-1} \in R_{i-1}} \left\{ \sum_{j=1}^{i-1} r_j : \mathcal{I}_{i-1}\big(\boldsymbol{r}_1^{i-1}\big) \geq s \right\}, \tag{25}$$

$$\mathcal{I}_{\min}(i) \equiv \min_{\boldsymbol{r}_1^i \in R_i} \mathcal{I}_i\big(\boldsymbol{r}_1^i\big), \qquad \mathcal{I}_{\max}(i) \equiv \max_{\boldsymbol{r}_1^i \in R_i} \mathcal{I}_i\big(\boldsymbol{r}_1^i\big). \tag{26}$$

As concrete examples of the last definition, for a generalized alpha investing rule, the current potentials $\{W(i)\}_{i\in\mathbb{N}}$, potentials at the last rejection $\{W(\tau_i)\}_{i\in\mathbb{N}}$ and total number of rejections $\{R(i)\}_{i\in\mathbb{N}}$ are index sequences.

THEOREM 3.6. *Consider a generalized alpha investing rule and assume that the test level* α_j *is determined based on index function* $\mathcal{I}_{j-1}$. *Namely, for each* $j \in \mathbb{N}$ *there exists a function* $g_j : \mathbb{R}_{\geq 0} \to [0,1]$ *such that* $\alpha_j = g_j(\mathcal{I}_{j-1}(\boldsymbol{R}_1^{j-1}))$. *Further, assume* $g_j(\cdot)$ *to be nondecreasing and weakly differentiable with weak derivative* $\dot{g}_j(s)$.

Then, the following upper bound holds for general dependencies among p-values:

$$\mathrm{FDR}(n) \leq \sum_{i=1}^{n} \left\{ g_i\big(\mathcal{I}_{\min}(i-1)\big) + \int_{I_{\min}(i-1)}^{I_{\max}(i-1)} \frac{\dot{g}_i(s)}{R_i^{\mathrm{L}}(s)+1}\,\mathrm{d}s \right\}. \tag{27}$$

The proof of this theorem is presented in the Supplementary Material [22].

EXAMPLE 3.7 (FDR control for dependent test statistics via modified LORD). We can modify LORD as to achieve FDR control even under dependent test statistics. As before, we let $\psi_i = b_0$. However, we fix a sequence $\boldsymbol{\xi} = (\xi_i)_{i\in\mathbb{N}}$, $\xi_i \geq 0$,

and set test levels according to rule $\alpha_i = \varphi_i = \xi_i W(\tau_i)$. In other words, compared with the original LORD procedure, we discount the capital accumulated at the last discovery as a function of the number of hypotheses tested so far, rather than the number of hypotheses tested since the last discovery.

This rule satisfies the assumptions of Theorem 3.6, with index sequence $\mathcal{I}_{i-1} = W(\tau_i)$ and $g_i(s) = \xi_i s$. Further, $\mathcal{I}_{\min}(0) = w_0$, $\mathcal{I}_{\min}(i-1) = b_0$ for $i \geq 2$, and $\mathcal{I}_{\max}(i-1) \leq w_0 + b_0(i-1)$, and $R_i^{\mathrm{L}}(s) \geq (\frac{s-w_0}{b_0})_+$. Substituting in equation (27) we obtain, assuming $w_0 \leq b_0$,

$$\begin{aligned}\mathrm{FDR}(n) &\leq w_0\xi_1 + \sum_{i=2}^{n}\Big(b_0\xi_i + \int_{b_0}^{w_0+b_0(i-1)} \frac{b_0\xi_i}{s - w_0 + b_0}\,\mathrm{d}s\Big)\\ &\leq w_0\xi_1 + \sum_{i=2}^{n} b_0\xi_i\big(1+\log(i)\big)\\ &\leq \sum_{i=1}^{n} b_0\xi_i\big(1+\log(i)\big).\end{aligned}$$

Hence, this rule controls FDR below level α under general dependency structure, if coefficients $(\xi_i)_{i\in\mathbb{N}}$ are set such that $\sum_{i=1}^{\infty}\xi_i(1+\log(i)) \leq \alpha/b_0$.

4. Statistical power. The class of generalized alpha investing rules is quite broad. In order to compare different approaches, it is important to estimate their statistical power.

Here, we consider a mixture model wherein each null hypothesis is false with probability π_1 independently of other hypotheses, and the p-values corresponding to different hypotheses are mutually independent. Under the null hypothesis H_i, we have p_i uniformly distributed in $[0,1]$ and under its alternative, p_i is generated according to a distribution whose c.d.f. is denoted by F. We let $G(x) = \pi_0 x + \pi_1 F(x)$, with $\pi_0 + \pi_1 = 1$, be the marginal distribution of the p-values. For presentation clarity, we assume that $F(x)$ is continuous.

While the mixture model is admittedly idealized, it offers a natural ground to compare online procedures to offline procedures. Indeed, online approaches are naturally favored if the true nonnulls arise at the beginning of the sequence of hypotheses, and naturally unfavored if they only appear later. On the other hand, if the p-values can be processed offline, we can always apply an online rule after a random re-ordering of the hypotheses. By exchangeability, we expect the performance to be similar to the ones in the mixture model.

The next theorem lower bounds the statistical power of LORD under the mixture model. This lower bound applies to any of the three versions of LORD.

THEOREM 4.1. *Consider the mixture model with $G(x)$ denoting the marginal distribution of p-values. Further, let $\Omega_0(n)$ [and its complement $\Omega_0^c(n)$] be the*

subset of true nulls (*nonnulls*), *among the first n hypotheses. Then, the average power of* LORD *rule is almost surely bounded as follows*:

$$\liminf_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i\in\Omega_0^c(n)} R_i \geq \left(\sum_{m=1}^{\infty} \prod_{\ell=1}^{m} (1 - G(b_0\gamma_\ell)) \right)^{-1}. \tag{28}$$

Proof of Theorem 4.1 is deferred to the Supplementary Material [22]. The lower bound is in fact the exact power for a slightly weaker rule that resets the potential at level b_0 after each discovery [in other words, equation (18) is replaced by $\varphi_i = \gamma_{i-\tau_i} b_0$]. This procedure is weaker only when multiple discoveries are made in a short interval of time. Hence, the above bound is expected to be accurate when π_1 is small, and discoveries are rare.

Recall that in LORD, parameters $\boldsymbol{\gamma} = (\gamma_\ell)_{\ell=1}^{\infty}$ can be any sequence of nonnegative, monotone nonincreasing numbers that sums up to one. This leaves a great extent of flexibility in choosing γ. The above lower bound on statistical power under the mixture model provides useful insight on what are good choices of $\boldsymbol{\gamma}$.

We first simplify the lower bound further. We notice that $\prod_{\ell=1}^{m}(1 - G(b_0\gamma_\ell)) \leq \exp(-\sum_{\ell=1}^{m} G(b_0\gamma_\ell))$. Further, by the monotonicity property of $\boldsymbol{\gamma}$, we have $G(b_0\gamma_\ell) \geq G(b_0\gamma_m)$ for $\ell \leq m$. Thus,

$$\lim_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i\in\Omega_0^c(n)} R_i \geq \mathcal{A}(G, \boldsymbol{\gamma}), \qquad \mathcal{A}(G, \boldsymbol{\gamma}) = \left(\sum_{m=1}^{\infty} e^{-mG(b_0\gamma_m)} \right)^{-1}.$$

In order to choose $\boldsymbol{\gamma}$, we use the lower bound $\mathcal{A}(G, \boldsymbol{\gamma})$ as a surrogate objective function. We let $\boldsymbol{\gamma}^{\mathrm{opt}}$ be the sequence that maximizes $\mathcal{A}(G, \boldsymbol{\gamma})$. The following proposition characterizes the asymptotic behavior of $\boldsymbol{\gamma}^{\mathrm{opt}}$.

PROPOSITION 4.2. *Let* $\boldsymbol{\gamma}^{\mathrm{opt}}$ *be the sequence that maximizes* $\mathcal{A}(G, \boldsymbol{\gamma})$ *under the constraint* $\sum_{\ell=1}^{\infty} \gamma_m = 1$. *Further suppose that* $F(x)$ *is concave and differentiable on an interval* $[0, x_0)$ *for some* $x_0 \in (0, 1)$. *Then there is a constant* $\eta = \eta(G, \pi_1)$ *independent of* m *such that, for all* m *large enough, the following holds true*:

$$\frac{1}{b_0} G^{-1}\left(\frac{1}{m}\log\left(\frac{m(1-\pi_1)}{\eta}\right)\right) \leq \gamma_m^{\mathrm{opt}} \leq \frac{1}{b_0} G^{-1}\left(\frac{2}{m}\log\left(\frac{1}{\eta G^{-1}(1/m)}\right)\right).$$

The proof of Proposition 4.2 is given in the Supplementary Material [22].

The concavity assumption of $F(x)$ requires the density of nonnull p-values [i.e., $F'(x)$] to be nonincreasing in a neighborhood $[0, x_0)$. This is a reasonable assumption because significant p-values are generically small and the assumption states that, in a neighborhood of zero, smaller values have higher density than larger values. In the Supplementary Material [22], we discuss compute the optimal sequence $\boldsymbol{\gamma}^{\mathrm{opt}}$ for two case examples.

5. Numerical simulations. In this section, we carry out some numerical experiments with synthetic data. For an application with real data, we refer to the Supplementary Material [22].

5.1. *Comparison with off-line rules.* In our first experiment, we consider hypotheses $\mathcal{H}(n) = (H_1, H_2, \dots, H_n)$ concerning the means of normal distributions. The null hypothesis is $H_j : \theta_j = 0$. We observe test statistics $Z_j = \theta_j + \varepsilon_j$, where ε_j are independent standard normal random variables. Therefore, one-sided p-values are given by $p_j = \Phi(-Z_j)$, and two sided p-values by $p_j = 2\Phi(-|Z_j|)$. Parameters θ_j are set according to a mixture model:

$$\theta_j \sim \begin{cases} 0, & \text{w.p. } 1 - \pi_1, \\ F_1, & \text{w.p. } \pi_1. \end{cases} \tag{29}$$

In our experiment, we set $n = 3000$ and and use the following three choices of the nonnull distribution:

Gaussian. In this case, the alternative F_1 is $\mathsf{N}(0, \sigma^2)$ with $\sigma^2 = 2\log n$. This choice of σ produces parameters θ_j in the interesting regime in which they are detectable, but not easily so. In order to see this recall that, under the global null hypothesis, $Z_i \sim \mathsf{N}(0, 1)$ and $\max_{i\in[n]} Z_i \sim \sqrt{2\log n}$ with high probability. Indeed $\sqrt{2\log n}$ is the minimax amplitude for estimation in the sparse Gaussian sequence model [8, 25].

In this case, we carry out two-sided hypothesis testing.

Exponential. In this case, the alternative F_1 is exponential $\mathsf{Exp}(\lambda)$ with mean $\lambda^{-1} = \sqrt{2\log n}$. The rationale for this choice is the same given above. The alternative is known to be nonnegative, and hence we carry out one-sided hypothesis testing.

Simple. In this example, the nonnulls are constant and equal to $A = \sqrt{\log n}$. Again, we carry out one-sided tests in this case.

We consider three online testing rules, namely alpha investing (AI), LORD (a special case of alpha spending with rewards) and Bonferroni. We also simulate the expected reward optimal (ERO) alpha investing rule introduced in [1]. For a brief overview of the ERO notion, recall that in a generalized alpha investing rule, pay-out φ_j, test level α_j and the reward ψ_j should satisfy inequalities (10) and (11). An ERO procedure finds the optimal point of trade-off between α_j and ψ_j, for a given value of φ_j, where optimality criterion is the expected reward of the current test, that is, $\mathbb{E}(R_j)\psi_j$. We compare performance of these online methods with the (adaptive) BH procedure, which as emphasized already, is an offline testing rule: it has access to the number of hypotheses and p-values in advance, while the former algorithms receive p-values in an online manner, without knowing the total number of hypotheses. We use Storey's variant of BH rule, that is better suited to cases in which the fraction of nonnulls π_1 is not necessarily small [36]. In all cases, we set as our objective to control FDR below $\alpha = 0.05$.

The different procedures are specified as follows.

Alpha investing. We set test levels according to

$$\alpha_j = \frac{W(j)}{1 + j - \tau_j}, \tag{30}$$

where τ_j denotes the time of the most recent discovery before time j. This proposal was introduced by [11] and boosts statistical power in cases in which the nonnull hypotheses appear in batches. We use parameters $w_0 = 0.005$ (for the initial potential), and $b_0 = \alpha - w_0 = 0.045$ (for the rewards). The rationale for this choice is that b_0 controls the evolution of the potential $W(n)$ for large n, while w_0 controls its initial value. Hence, the behavior of the resting rule for large n is mainly driven by b_0.

Note that by [1], Corollary 2, this is an ERO alpha investing rule,[7] under the Gaussian and exponential alternatives. It is worth noting that for the case of exponential alternatives, alpha investing is indeed an ERO procedure, cf. [1], Theorem 2.

ERO alpha investing. For the case of simple alternative, the maximum power achievable at test i is $\rho_i = \Phi(A + \Phi^{-1}(\alpha_i))$. In this case, we consider ERO alpha investing [1] defined by $\varphi_i = (1/10) \cdot W(i-1)$, and with α_i, ψ_i given implicitly by the solution of $\varphi_i/\rho_i = \varphi_i/\alpha_i - 1$ and $\psi_i = \varphi_i/\alpha_i + b_0 - 1$. We use parameters $b_0 = 0.045$ and $w_0 = 0.005$.

LORD. We use LORD 3 and choose the sequence $\boldsymbol{\gamma} = (\gamma_m)_{m \in \mathbb{N}}$ as follows:

$$\gamma_m = C \frac{\log(m \vee 2)}{m e^{\sqrt{\log m}}}, \tag{31}$$

with C determined by the condition $\sum_{m=1}^{\infty} \gamma_m = 1$, which yields $C \approx 0.07720838$. This choice of $\boldsymbol{\gamma}$ is loosely motivated by Example E.2, given in the Supplementary Material [22]. Notice, however, that we do not assume the data to be generated with the model treated in that example. Further, for this case we set parameters $w_0 = 0.005$ (for the initial potential), and $b_0 = 0.045$ (for the rewards).

Bonferroni. We set the test levels as $\alpha_m = \gamma_m \alpha$, where the values of γ_m are set as per equation (31), and therefore $\sum_{m=1}^{\infty} \alpha_m = \alpha$.

Storey. It is well known that the classical BH procedure satisfies FDR $\leq \pi_0 \alpha$ where π_0 is the proportion of true nulls. A number of adaptive rules have been proposed that use a plug-in estimate of π_0 as a multiplicative correction in the BH procedure [23, 24, 30, 36]. Following [6], the adaptive test thresholds are given by $\alpha H(\boldsymbol{p}) i/n$ (instead of $\alpha i/n$), where $H(\boldsymbol{p})$ is an estimate of π_0^{-1}, determined as a function of p-values, $\boldsymbol{p} = (p_1, \ldots, p_n)$.

Here, we focus on Storey-λ estimator given by [36]:

$$H(\boldsymbol{p}) = \frac{(1-\lambda)n}{\sum_{i=1}^{n} \mathbb{I}(p_i > \lambda) + 1}. \tag{32}$$

[7] Note that, since θ_j is unbounded under the alternative the maximal power is equal to one.

Storey's estimator is in general an underestimate of π_0^{-1}. A standard choice of $\lambda = 1/2$ is used in the SAM software [37]. In [6], it is shown that the choice $\lambda = \alpha$ can have better properties under dependent p-values. In our simulations, we tried both choices of λ.

Our empirical results are presented in Figure 1. As we see, all the rules control FDR below the nominal level $\alpha = 0.05$, as guaranteed by Theorem 3.1. While BH and the generalized alpha investing schemes (LORD, alpha investing, ERO alpha investing) exploit most of the allowed amount of false discoveries, Bonferroni is clearly too conservative. A closer look reveals that the generalized alpha investing schemes are somewhat more conservative than BH. Note however that the present simulations assume the nonnulls to arrive at random times, which is a more benign scenario than the one considered in Theorem 3.1, where arrival times of nonnulls are adversarial.

In terms of power, LORD appears particularly effective for small π_1, while standard alpha investing suffers a loss of power for large π_1. This is related to the fact that $\varphi_j = \alpha_j/(1-\alpha_j)$ in this case. As a consequence, the rule can effectively stop after a large number of discoveries, because α_j gets close to one.

Figure 2 showcases the FDR achieved by various rules as a function of α, for $\pi_1 = 0.2$ and exponential alternatives. For alpha investing and LORD, we use parameters $b_0 = 0.9\alpha$ and $w_0 = 0.1\alpha$. The generalized alpha investing rules under consideration have FDR below the nominal α, and track it fairly closely. The gap is partly due to the fact that, for large number of discoveries, the FDR of generalized alpha investing rules is closer to b_0 than to $\alpha = b_0 + w_0$, cf. Remark 3.2.

5.2. *The effect of ordering*. By definition, the BH rule is insensitive to the order in which the hypotheses are presented. On the contrary, the outcome of online testing rules depends on this ordering. This is a weakness, because the ordering of hypotheses can be adversarial, leading to a loss of power, but also a strength. Indeed, in some applications, hypotheses can be ordered, using side information, such that those most *likely* to be rejected come first. In these cases, we expect generalized alpha investing procedures to be potentially *more powerful than benchmark offline rules* as BH.

For instance, Li and Barber [27] analyze a drug-response dataset proceeding in two steps. First, a family of hypotheses (gene expression levels) are ordered using side information, and then a multiple hypothesis testing procedure is applied to the ordered data.[8] Other approaches, such as distributing the weights unevenly among the hypotheses [13] are also potentially useful in settings where there is side information about the hypotheses that are more likely to be nonnull.

[8]The procedure of [27] is designed as to reject the first $\hat{k}$ null hypotheses, and accept the remaining $n - \hat{k}$. However, this specific structure is a design choice, and is not a constraint arising from the application.

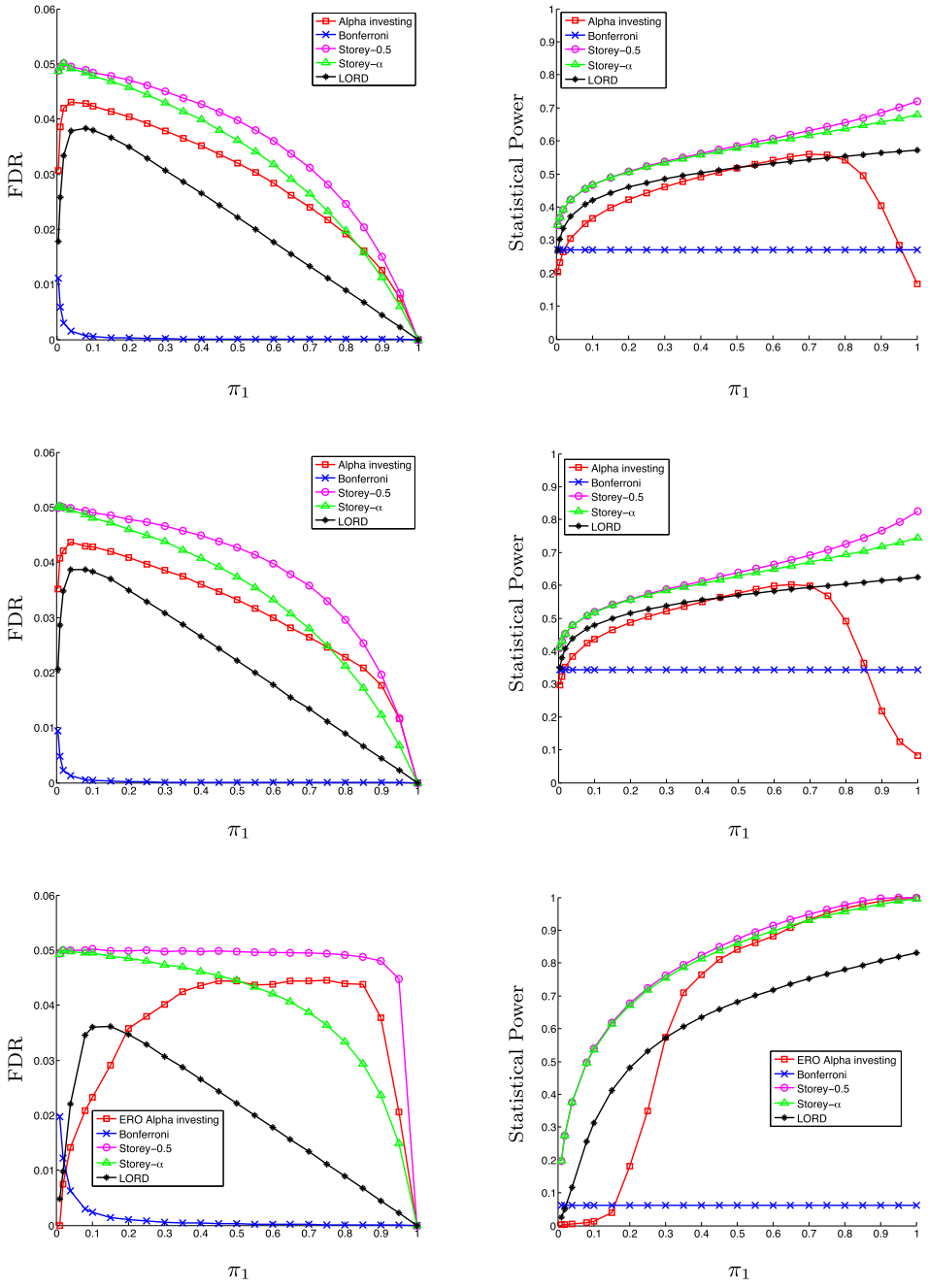


FIG. 1. *FDR and statistical power versus fraction of nonnull hypotheses* π_1 *for setup described in Section 5 with* $\alpha = 0.05$. *The three rows correspond to Gaussian, exponential, and simple alternatives (from top to bottom). FDR and power are computed by averaging over* 20,000 *independent trials (for Gaussian and exponential alternatives) or* 500 *trials (for simple alternatives). Here hypotheses are considered in random order of arrival.*

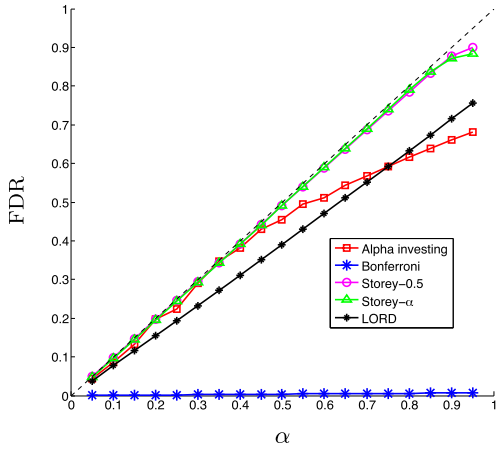


FIG. 2. *FDR achieved by various methods compared to the target FDR* α *as* α *varies. Here we use* $n = 3000$ *hypotheses with a proportion* $\pi_1 = 0.2$ *of nonnulls and exponential alternatives. The FDR is estimated by averaging over* 20,000 *independent trials.*

In order to explore the effect of a favorable ordering of the hypotheses, we reconsider the exponential model in the previous section, and simulate a case in which side information is available. For each trial, we generate the mean $(\theta_j)_{1\le j\le n}$, and two independent sets of observations $Z_j = \theta_j + \varepsilon_j$, $Z'_j = \theta_j + \varepsilon'_j$, with $\varepsilon_j \sim \mathsf{N}(0,1)$, $\varepsilon'_j \sim \mathsf{N}(0,\sigma^2)$ independent. We then compute the corresponding (one-sided) p-values $(p_j)_{1\le j\le n}$, $(p'_j)_{1\le j\le n}$. We use the p-values $(p'_j)_{1\le j\le n}$ to order the hypotheses[9] (in such a way that these p-values are increasing along the ordering). We then use the other set of p-values $(p_j)_{1\le j\le n}$ to test the null hypotheses $H_{j,0} : \theta_j = 0$ along this ordering.

Let us emphasize that, for this simulation, better statistical power would be achieved if we computed a single p-value p_j by processing jointly Z_j and Z'_j. However, in real applications, the two sources of information are heterogenous and this joint processing is not warranted, see [27] for a discussion of this point.

Figure 3 reports the FDR and statistical power in this setting. We used LORD with parameters $(\gamma_m)_{m\ge1}$ given by equation (31), and simulated two noise levels for the side information: $\sigma^2 = 1$ (noisy ordering information) and $\sigma^2 = 1/2$ (less noisy ordering). As expected, with a favorable ordering the FDR decreases significantly. The statistical power increases as long as the fraction of nonnulls π_1 is not too large. This is expected: when the fraction of nonnulls is large, ordering is less relevant.

[9]Note that ordering by increasing p'_j is equivalent to ordering by decreasing $|Z'_j|$ and the latter can be done without knowledge of the noise variance σ^2.

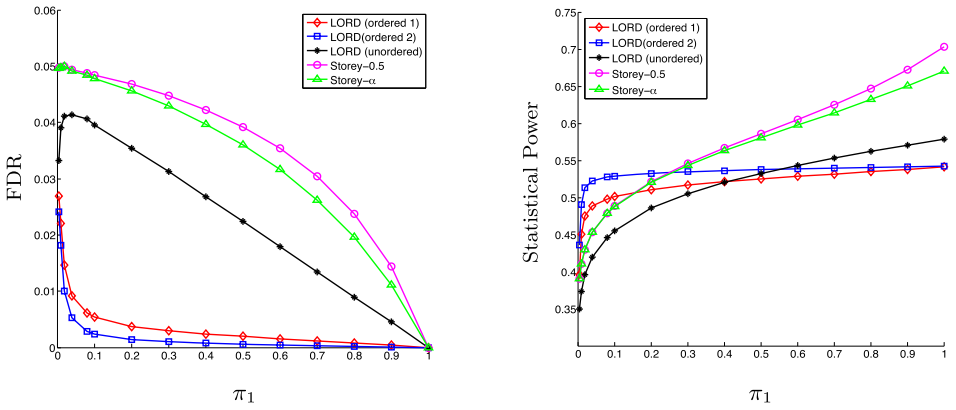


FIG. 3. *FDR and statistical power for* LORD *with favorably ordered hypotheses* (*setup of Section* 5.2). *Here* $n = 3000$, $\pi_1 = 0.2$ *and data are obtained by averaging over* 20,000 *trials. Unordered*: *Null and nonnull hypotheses are ordered at random. Ordered* 1: *hypotheses are ordered using very noisy side information* ($\sigma^2 = 1$). *Ordered* 2: *hypotheses are ordered using less noisy side information* ($\sigma^2 = 1/2$).

In particular, for small π_1, the gain in power can be as large as 20% (for $\sigma^2 = 1$) and as 30% (for $\sigma^2 = 1/2$). The resulting power is superior to adaptive BH [36] for $\pi_1 \gtrsim 0.15$ (for $\sigma^2 = 1$), or $\pi_1 \gtrsim 0.25$ (for $\sigma^2 = 1/2$).

5.3. FDR *control versus* mFDR *control.* Aharoni and Rosset [1] proved that generalized alpha investing rules control mFDR_{w_0/b_0}. Formally,

$$\mathrm{mFDR}_{w_0/b_0}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \frac{\mathbb{E}V^{\theta}(n)}{\mathbb{E}R(n) + (w_0/b_0)} \le b_0. \tag{33}$$

As mentioned before (see also the Supplementary Material [22]), this metric has been criticized because it does not control a property of the realized sequence of tests; instead it controls a ratio of expectations.

Our Theorem 3.3 controls a different metric that we called $\mathrm{sFDR}_{\eta}(n)$:

$$\mathrm{sFDR}_{w_0/b_0}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left\{\frac{V^{\theta}(n)}{R(n) + (w_0/b_0)}\right\} \le b_0. \tag{34}$$

This quantity is the expected ratio, and hence passes the above criticism. Note that both theorems yield control at level $\alpha = b_0$, for the same class of rules.

Finally, Theorem 3.1 controls a more universally accepted metric, namely FDR, at level $\alpha = w_0 + b_0$. A natural question is whether, in practice, we should choose w_0, b_0 as to guarantee FDR control (and hence set $w_0 + b_0 \le \alpha$) or instead be satisfied with mFDR and sFDR control, which allow for $b_0 = \alpha$ and hence potentially larger statistical power.

While an exhaustive answer to this question is beyond the scope of this paper, we repeated the simulations in Figure 1, using the two different criteria. The results, provided in the Supplementary Material [22], suggest that this question might not have a simple answer. On one hand, under the setting of Figure 1 (independent p-values, large number of discovery) mFDR and sFDR seem stringent enough criteria. On the other, the gain in statistical power that is obtained from these criteria, rather than FDR, is somewhat marginal.

6. Control of false discovery exceedance. Ideally, we would like to control the proportion of false discoveries in any given realization of our testing procedures. We recall that this is given by [cf. equation (20)]

$$\mathrm{FDP}^{\theta}(n) \equiv \frac{V^{\theta}(n)}{R(n) \vee 1}. \tag{35}$$

False discovery rate is the *expected* proportion of false discoveries. However—in general—control of FDR does not prevent FDP from varying, even when its average is bounded. In real applications, the actual FDP might be far from its expectation. For instance, as pointed out by Owen [31], the variance of FDP can be large if the test statistics are correlated.

Motivated by this concern, the *false discovery exceedance* is defined as

$$\mathrm{FDX}_{\gamma}(n) \equiv \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{P}\big(\mathrm{FDP}^{\theta}(n) \geq \gamma\big) \tag{36}$$

for a given tolerance parameter $\gamma \geq 0$. Controlling FDX instead of FDR gives a stronger preclusion from large fractions of false discoveries.

Several methods have been proposed to control FDX in an offline setting. Van der Laan, Dudoit and Pollard [40] observed that any procedure that controls FWER, if augmented by a sufficiently small number of rejections, also controls FDX. Genovese and Wasserman [14] suggest controlling FDX by inverting a set of uniformity tests on the vector of p-values. Lehmann and Romano [26] proposed a step-down method to control FDX.

A natural criterion to impose in the online setting would be the control of $\sup_{n \geq 1} \mathrm{FDX}_{\gamma}(n)$. However, this does not preclude the possibility of large proportions of false discoveries at some (rare) random times n. It could be—as a cartoon example—that $\mathrm{FDP}^{\theta}(n) = 1/2$ independently with probability α at each n, and $\mathrm{FDP}^{\theta}(n) = \gamma/2$ with probability $1 - \alpha$. In this case, $\sup_{n \geq 1} \mathrm{FDX}_{\gamma}(n) \leq \alpha$ but $\mathrm{FDP}^{\theta}(n) = 1/2$ almost surely for infinitely many times n. This is an undesirable situation.

A more faithful generalization of FDX to the online setting is therefore

$$\mathrm{FDX}_{\gamma} \equiv \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{P}\Big(\sup_{n \geq 1} \mathrm{FDP}^{\theta}(n) \geq \gamma\Big). \tag{37}$$

We will next propose a class of generalized alpha investing rules for online control of FDX_{γ}.

6.1. *The effect of reducing test levels.* Before describing our approach, we demonstrate through an example that the FDP can differ substantially from its expectation. We also want to illustrate how a naive modification of the previous rules only achieves a better control of this variability at the price of a significant loss in power.

Note that the desired bound $\mathrm{FDP}^{\theta}(n) < \gamma$ follows if we can establish $b_0 R(n) - V(n) + (\gamma - b_0) > 0$ for some $\gamma \geq b_0 \geq 0$. Recall that a generalized alpha investing procedure continues until the potential $W(n)$ remains nonnegative. Therefore, for such a procedure, it suffices to bound the probability that the stochastic process $B(n) \equiv b_0 R(n) - W(n) - V(n) + (\gamma - b_0)$ crosses zero. As we show in Lemma F.1, $B(n)$ is a submartingale, and thus in expectation it moves away from zero. In order to bound the deviations from the expectation, consider the submartingale increments $B_j \equiv B(j) - B(j-1)$ given by

$$B_j = (b_0 - \psi_j) R_j + \varphi_j - V_j. \tag{38}$$

If the jth null hypothesis is false, that is, $\theta_j \neq 0$, we have $V_j = 0$ and $B_j \geq 0$ by invoking assumption G1 and noting that $R_j \in \{0, 1\}$. Under the null hypothesis, $V_j = R_j$, and

$$\mathrm{Var}(B_j | \mathcal{F}_{j-1}) = (b_0 - \psi_j - 1)^2 \alpha_j (1 - \alpha_j). \tag{39}$$

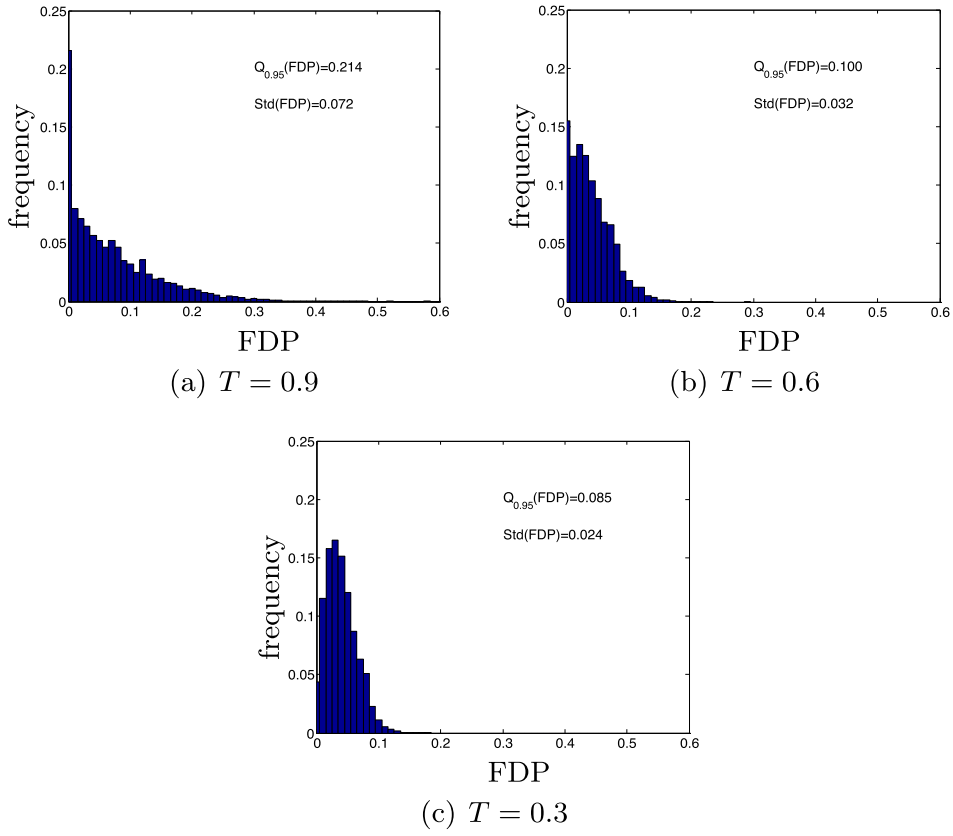
Reducing $\mathrm{Var}(B_j | \mathcal{F}_{j-1})$ lowers variations of the submartingale and hence the variation of the false discovery proportions. Note that for a generalized alpha investing rule, if we keep b_0, ψ_j unchanged and lower the test levels α_j, the rule still satisfies conditions G1, G2 and thus controls FDR at the desired level. On the other hand, this modification decreases $\mathrm{Var}(B_j | \mathcal{F}_{j-1})$ as per equation (39). In summary, reducing the test levels has the effect of reducing the variation of false discovery proportion at the expense of reducing statistical power.

We carry out a numerical experiment within a similar setup as the one discussed in Section 5. A set of n hypotheses are tested, each specifying mean of a normal distribution, $H_j : \theta_j = 0$. The test statistics are independent, normally distributed random variables $Z_j \sim \mathsf{N}(\theta_j, 1)$. For nonnull hypotheses, we set $\theta_j = 3$. The total number of tests is $n = 1000$ of which the first 100 are nonnull.

We consider three different testing rules, namely alpha investing, alpha spending with rewards and LORD, all ensuring FDR control at level $\alpha = 0.05$. The details of these rules as well as the choice of parameters is the same as Section 5.

In order to study the effect of reducing test levels, for each of these rules we truncate them by a threshold value T, that is, we use $\alpha_j^T = \alpha_j \vee T$. We plot the histogram of false discovery proportions using 30,000 replications of the test statistics sequence. We further report standard deviation and 0.95 quantile of FDPs. The results are shown in Figures 4, 5, 6.

As a first remark, while all of the rules considered control FDR below $\alpha = 0.05$, the actual false discovery proportion in Figures 4, 5, 6 has a very broad



(a) $T = 0.9$ (b) $T = 0.6$ (c) $T = 0.3$

FIG. 4. *Histogram of FDP for alpha investing rule with different values of* T.

distribution. Consider for instance alpha investing, at threshold level $T = 0.9$. Then FDP exceeds 0.15 (three times the nominal value) with probability 0.13.

Next, we notice that reducing the test levels (by reducing T) has the desired effect of reducing the variance of the FDP. This effect is more pronounced for alpha investing. Nevertheless quantifying this effect is challenging due to the complex dependence between B_j and history $\mathcal{F}_{j-1}$. This makes it highly nontrivial to adjust threshold T to obtain $\mathrm{FDX}_\gamma \le \alpha$. In the next section, we achieve this through a different approach.

6.2. *Rules for controlling* FDX_γ. Let $M(0) = \gamma - b_0 - w_0 > 0$ and define, for $n \in \mathbb{N}$, $M(n) = M(0) + \sum_{j=1}^{n} M_j$, where

$$M_j \equiv \max\{(1 + \psi_j - b_0)(\alpha_j - R_j), (b_0 - \psi_j)R_j, \psi_j - b_0\}. \tag{40}$$

Note that $M(n)$ is a function of $(R_1, \dots, R_n)$, that is, it is measurable on $\mathcal{F}_n$. We then require the following conditions in addition to G1 and G2 introduced in Section 2.1:

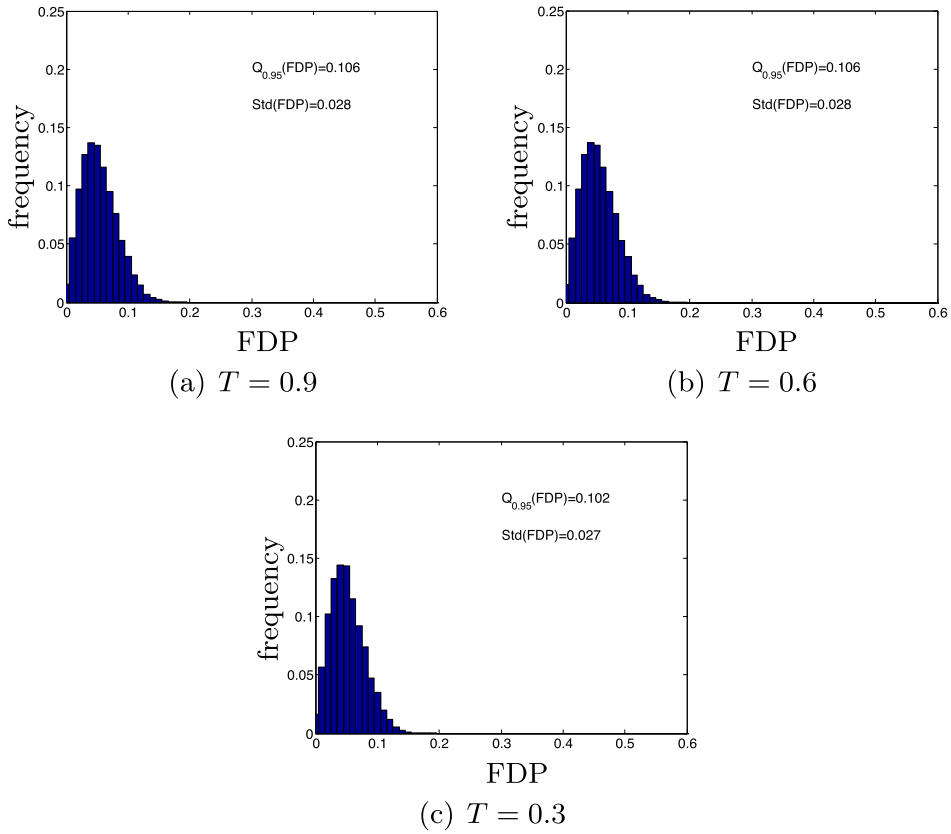


FIG. 5. *Histogram of FDP for alpha spending with rewards for different values of* T.

G3. $w_0 < \gamma - b_0$.

G4. For $j \in \mathbb{N}$ and all $\boldsymbol{R}_1^j \in \{0,1\}^j$, if

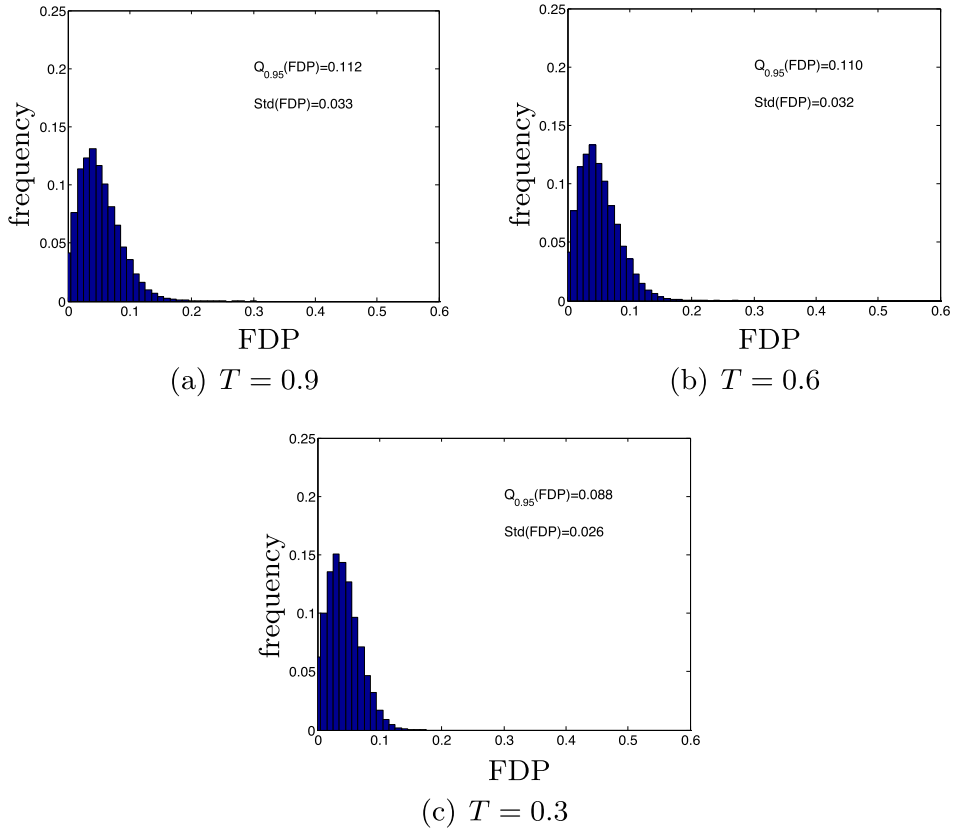
$$M(j) + \xi_{j+1} > \frac{\gamma - b_0 - w_0}{1-\alpha}, \tag{41}$$

then $\alpha_i = 0$ for all $i > j$, where we define $\xi_j \equiv \max\{(1 + \psi_j - b_0)\alpha_j, |b_0 - \psi_j|\}$.

Condition G4 is well posed since $M(j)$ and ξ_{j+1} are functions of $\boldsymbol{R}_1^j$.

Note that any generalized alpha investing rule can be modified as to satisfy these conditions. Specifically, the rule keeps track of LHS of (41) (it is an *observable* quantity) and whenever inequality (41) is violated, the test levels are set to zero onwards, that is, $\alpha_i = 0$ for $i \geq j$. The sequence $(\xi_j)_{j\in\mathbb{N}}$ is constructed in a way to be a predictable process that bounds M_j. Consequently, $M(j) + \xi_{j+1} \in \mathcal{F}_j$ bounds $M(j+1)$.

The decrement and increment values φ_j and ψ_j are determined in way to satisfy conditions G2 and G5.



(a) $T = 0.9$ (b) $T = 0.6$

(c) $T = 0.3$

FIG. 6. *Histogram of FDP for LORD rule with different values of* T.

We then establish FDX control under a certain negative dependency condition on the test statistics.

THEOREM 6.1. *Assume that the* p*-values* $(p_i)_{i\in\mathbb{N}}$ *are such that, for each* $j \in \mathbb{N}$, *and all* $\boldsymbol{\theta} \in H_j$ (*i.e., all* $\boldsymbol{\theta}$ *such that the null hypothesis* $\theta_j = 0$ *holds*), *we have*

$$\mathbb{P}_{\boldsymbol{\theta}}(p_j \le \alpha_j | \mathcal{F}_{j-1}) \le \alpha_j, \tag{42}$$

almost surely.

Then, any generalized alpha investing rule that satisfies conditions G3, G4 *above* (*together with* G1 *and* G2) *controls the false discovery exceedance*:

$$\mathrm{FDX}_\gamma \le \alpha. \tag{43}$$

The proof of this theorem is presented in the Supplementary Material [22]. Notice that the dependency condition (42) is satisfied, in particular, if the p-values are independent.

TABLE 1
FDX_γ and FDR *for* LORD *with stopping criterion* (44) *using* 30,000 *realizations of the test statistics. Here,* $\alpha = 0.05$ *and* $\gamma = 0.15$, *and* π_1 *represents the fraction of truly nonnull hypotheses that appear at the beginning of the stream as described in Section* 6.1

Online control of FDX_γ using stopping criterion (44)					
π_1	0.005	0.01	0.02	0.03	0.04
FDX_γ	0.028	0.004	0.000	0.000	0.000
FDR	0.006	0.005	0.005	0.005	0.005
Power	0.666	0.699	0.679	0.658	0.639

EXAMPLE 6.2. For given values of $\alpha \in (0, 1)$ and $\gamma \in (\alpha, 1)$, consider LORD algorithm with $b_0 = \alpha$, $\psi_j = \alpha$ for $j \in \mathbb{N}$ and $w_0 = (\gamma - \alpha)/2$. By equation (40) we have $M_j = \alpha_j \mathbb{I}(R_j = 0)$. In order to satisfy condition G4, the rule keeps track of $M(n)$ and stops as soon as inequality (41) is violated:

$$\alpha_{n+1} + \sum_{i=1}^{n} \alpha_i \mathbb{I}(R_i = 0) > \frac{\gamma - \alpha}{2(1 - \alpha)}. \tag{44}$$

Note that for LORD, the potential sequence $W(n)$ always remain positive and thus the stopping criterion is defined solely based on the above inequality. Clearly, this rule satisfies assumptions G1, G2, G3, G4 and by applying Theorem 6.1 ensures $\mathrm{FDX}_\gamma \le \alpha$.

We use the above rule to control false discovery exceedance for the simulation setup described in Section 6.1 for values of $\alpha = 0.05$ and $\gamma = 0.15$. The results are summarized in Table 1. The false discovery rates and proportions are estimated using 30,000 realizations of test statistics. As we see the rule controls both FDR and FDX_γ below α.

7. Discussion. Our main result is that all generalized alpha investing rules control FDR, provided they satisfy a natural monotonicity condition. This result can be regarded as reinforcing and complementing the conclusions of [1] which introduced generalized alpha investing, and proved mFDR control. Since the two metrics can be significantly different, with FDR somewhat more broadly accepted, this should develop more confidence towards the practical use of these methods.

Within this broad family, we believe that LORD is mainly appealing because of its simplicity: testing levels only depend on the the time of the most recent discovery, and not on the whole past. This property also simplifies the analysis of LORD. In particular, in Section 4 we obtained bounds on the statistical power of the LORD under the mixture model, that could be used to set the parameters of the rule. Further, a simple modification of LORD was suggested for the case of dependent p-values; cf. Section 3.2.

While our work broadly supports the use of generalized alpha investing rules (and, in particular, LORD), we believe that extra caution should be taken when the false discovery proportion can deviate significantly from its expectation (which is the FDR). This can be the case when the number of hypotheses is not very large, or there is significant correlation. In this case, the false discovery exceedance (FDX) is a more meaningful metric, and additional constraints should be imposed on generalized alpha investing rules.

Acknowledgments. The authors would like to thank the Co-Editor, Associate Editor and referees for their valuable comments that helped us improve the paper significantly.

SUPPLEMENTARY MATERIAL

Online rules for control of false discovery rate and false discovery exceedance (DOI: 10.1214/17-AOS1559SUPP; .pdf). Due to space constraints, proof of theorems and some of the technical details as well as additional numerical studies are provided in the Supplementary Material [22].

REFERENCES

[1] AHARONI, E. and ROSSET, S. (2014). Generalized α-investing: Definitions, optimality results and application to public databases. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 771–794. MR3248676

[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876

[3] BARBER, R. F. and CANDES, E. J. (2016). A knockoff filter for high-dimensional selective inference. Available at arXiv:1602.03574.

[4] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

[5] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[6] BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

[7] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[8] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886

[9] DWORK, C., FELDMAN, V., HARDT, M., PITASSI, T., REINGOLD, O. and ROTH, A. (2015). Preserving statistical validity in adaptive data analysis [extended abstract]. In *STOC'*15—*Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 117–126. ACM, New York. MR3388189

[10] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Available at arXiv:1410.2597.

[11] FOSTER, D. P. and STINE, R. A. (2008). α-investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 429–444. MR2424761

[12] GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15** 870–878.

[13] GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

[14] GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468

[15] G'SELL, M. G., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 423–444. MR3454203

[16] IOANNIDIS, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Jornal of the American Medical Association* **294** 218–228.

[17] IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *Chance* **18** 40–47. MR2216666

[18] JAVANMARD, A. and MONTANARI, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In 51*st Annual Allerton Conference* 1427–1434, Monticello, IL.

[19] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[20] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

[21] JAVANMARD, A. and MONTANARI, A. (2015). On online control of false discovery rate. Available at arXiv:1502.06197.

[22] JAVANMARD, A. and MONTANARI, A. (2018). Supplement to "Online rules for control of false discovery rate and false discovery exceedance." DOI:10.1214/17-AOS1559SUPP.

[23] JIN, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 461–493. MR2420411

[24] JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

[25] JOHNSTONE, I. M. (1994). On minimax estimation of a sparse normal mean vector. *Ann. Statist.* **22** 271–289. MR1272083

[26] LEHMANN, E. L. and ROMANO, J. P. (2012). *Generalizations of the Familywise Error Rate*. Springer, Berlin.

[27] LI, A. and BARBER, R. F. (2016). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Statist. Assoc.* **112** 1–38. MR3671774

[28] LIN, D., FOSTER, D. P. and UNGAR, L. H. (2011). VIF regression: A fast regression algorithm for large data. *J. Amer. Statist. Assoc.* **106** 232–247. MR2816717

[29] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

[30] MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. MR2275246

[31] OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 411–426. MR2155346

[32] PEKELIS, L., WALSH, D. and JOHARI, R. (2015). The new stats engine. Available at http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf.

[33] PRINZ, F., SCHLANGE, T. and ASADULLAH, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* **10** 712–712.

[34] REINER, A., YEKUTIELI, D. and BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19** 368–375.

[35] ROSSET, S., AHARONI, E. and NEUVIRTH, H. (2014). Novel statistical tools for management of public databases facilitate community-wide replicability and control of false discovery. *Genetic Epidemiology* **38** 477–481.

[36] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302

[37] STOREY, J. D. and TIBSHIRANI, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data* 272–290. Springer, New York. MR2001400

[38] TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689

[39] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[40] VAN DER LAAN, M. J., DUDOIT, S. and POLLARD, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 15, 27. MR2101464

[41] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

DATA SCIENCES AND OPERATIONS DEPARTMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: ajavanma@usc.edu

DEPARTMENT OF ELECTRICAL ENGINEERING
AND DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: montanar@stanford.edu