Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(4) 976–993

DOI: 10.1177/0962280220983381
journals.sagepub.com/home/smm
SAGE

# Online control of the familywise error rate

**Jinjin Tian**[1] and **Aaditya Ramdas**[1,2]

**Abstract**

Biological research often involves testing a growing number of null hypotheses as new data are accumulated over time. We study the problem of online control of the familywise error rate, that is testing an a priori unbounded sequence of hypotheses ($p$-values) one by one over time without knowing the future, such that with high probability there are no false discoveries in the entire sequence. This paper unifies algorithmic concepts developed for offline (single batch) familywise error rate control and online false discovery rate control to develop novel online familywise error rate control methods. Though many offline familywise error rate methods (e.g., Bonferroni, fallback procedures and Sidak's method) can trivially be extended to the online setting, our main contribution is the design of new, powerful, adaptive online algorithms that control the familywise error rate when the $p$-values are independent or locally dependent in time. Our numerical experiments demonstrate substantial gains in power, that are also formally proved in an idealized Gaussian sequence model. A promising application to the International Mouse Phenotyping Consortium is described.

**Keywords**

Online multiple testing, FWER control, null proportion adaptivity, discarding conservative nulls

## 1 Introduction

Modern genomics studies typically require large scale multiple hypotheses testing, which are sometimes conducted in an online manner (meaning one or few hypotheses at a time), not as a big large batch of hypotheses tested all at once. Thus, the family of tested hypotheses is continually growing over time, due to the accumulation of data (both types and amounts). For example, one international scientific project, the International Mouse Phenotyping Consortium (IMPC), aims to create and characterize the phenotype of 20,000 knockout mouse strains; this was launched in September 2011, and was projected to take 10 years. Available datasets and the resulting family of hypotheses constantly grow as new knockouts are studied, while discovery-inspired downstream analyses are conducted along the way. Thus we are faced with a nonstandard multiple hypothesis testing problem, one in which the nature or number of hypotheses (or $p$-values) is not known in advance but become known one at a time. This exemplifies the challenge of online hypothesis testing, where at each step the scientist must decide whether or not to reject the current null hypothesis without knowing the future hypotheses or their outcomes (rejected or not).

Formally, online multiple testing refers to the setting in which a potentially infinite stream of hypotheses $H_1, H_2, \ldots$ (respectively $p$-values $P_1, P_2, \ldots$) is tested one by one over time. It is important to distinguish the aforementioned setup from sequential hypothesis testing, where a single hypothesis is repeatedly tested as new data are collected. Our setup zooms out one level, and it is the hypotheses that appear one at a time, each one summarized by a $p$-value.

[1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA

**Corresponding author:**
Jinjin Tian, Department of Statistics and Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh 15213-3815, PA, USA.
Email: jinjint@andrew.cmu.edu

At each step $t \in \mathrm{N}$, one must decide whether to reject the current null hypothesis $H_t$ or not, without knowing the outcomes of all the future tests. Typically, we reject the null hypothesis when $P_t$ is smaller than some threshold $\alpha_t$. Let $\mathcal{R}(T)$ represent the set of rejected null hypotheses by time $T$, and *nlls* be the unknown set of true null hypotheses; then, $\mathcal{V}(T) \equiv \mathcal{R}(T) \cap \mathcal{H}_0$ is the set of incorrectly rejected null hypotheses, also known as false discoveries. Denoting $V(T) = |\mathcal{V}(T)|$, some error metrics are the false discovery rate (FDR), familywise error rate (FWER) and power which are defined as

$$\begin{aligned} &\mathrm{FDR}(T) \equiv \mathbb{E}\left[\frac{V(T)}{|\mathcal{R}(T)| \vee 1}\right], \quad \mathrm{FWER}(T) \equiv \Pr\left\{V(T) \geq 1\right\}, \\ &\mathrm{power}(T) \equiv \mathbb{E}\left[\frac{|\mathcal{H}_0^c \cap \mathcal{R}(T)|}{|\mathcal{H}_0^c|}\right] \end{aligned} \tag{1}$$

It is arguably of interest to control the FDR or FWER at each time below a prespecified level $\alpha \in (0, 1)$, while maximizing power. Such a goal is also essential with respect to providing a reliable resource for downstream analysis along the way. Notice that $\mathrm{FWER}(T)$ is monotonically nondecreasing in T, therefore controlling FWER $(T)$ at each $T \in \mathrm{N}$ is equivalent to controlling $\lim_{T\to\infty} \mathrm{FWER}(T)$. Therefore, we drop the index $T$, and only discuss techniques controlling $\lim_{T\to\infty} \mathrm{FWER}(T)$.

There is a large variety of procedures for (A) offline (single batch) FDR control,[1–3] (B) online FDR control,[4–9] and (C) offline FWER control.[10–12] However, the online FWER problem is underexplored; for example, in their seminal online FDR paper, Dean Foster and Robert Stine[4] only mention in passing that for online FWER control, one can simply use an online Bonferroni method (that they term Alpha-Spending). We first point out that offline FWER procedures can be easily extended to have online counterparts (like Sidak's and fallback methods). However, our main contribution is to derive new "adaptive" procedures for online FWER control that are demonstrably more powerful under independence or local (in time) dependence assumptions.

In experiments, we find that our online extensions of classical offline FWER control methods like Sidak[13] and fallback[12,14] have a fairly negligible power improvement over Alpha-Spending when non-nulls are interspersed randomly over time. Also, as presented in Figure 1, these algorithms are sub-optimal when they encounter a non-vanishing proportion of non-nulls, or a significant proportion of *conservative* null*.[a] In contrast, our adaptive discarding (ADDIS) algorithms are more powerful online FWER control methods that adapt to both an unknown fraction of non-nulls and unknown conservativeness of nulls (Section 3.3 and Figure 2). As shown in Figure 1, ADDIS-Spending is significantly more powerful than Alpha-Spending, online Sidak, online fallback, etc. We are able to theoretically justify this power superiority in an idealized Gaussian model (Section 4), and also confirm it in the real IMPC dataset we mentioned earlier (Section 5.2). Although these powerful new adaptive methods require independence to have valid FWER control, they can be easily generalized to handle "local dependence", an arguably more realistic dependence structure in practice (Section 3.4).

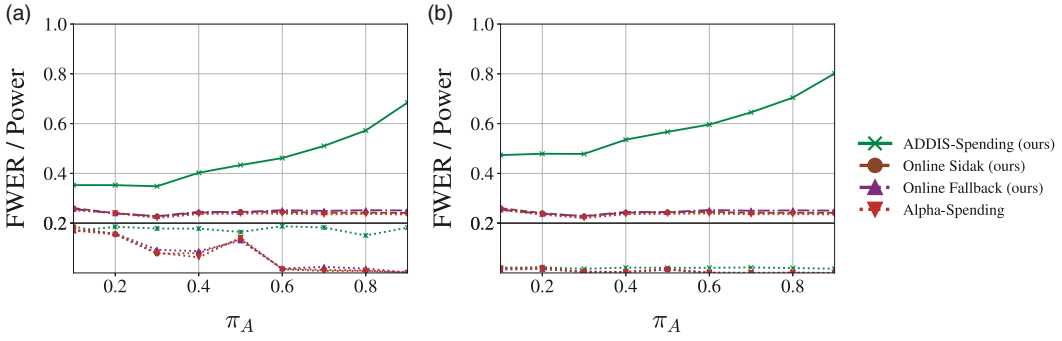


**Figure 1.** Statistical power and FWER versus fraction of non-null hypotheses $\pi_A$ for Alpha-Spending (online Bonferroni) and our new algorithms (ADDIS-Spending, Online Fallback and Online Sidak) at target FWER level $\alpha = 0.2$ (solid black line). The curves above the horizontal line at 0.2 display the achieved power of each methods versus $\pi_A$, while the lines below 0.2 display the achieved FWER of each methods versus $\pi_A$. The experimental setting is described in Section 5.1: all observations are Gaussians, we set the alternative mean $\mu_A = 4$ for both figures, but we set the null mean $\mu_N = 0$ for the left figure and $\mu_N = -1$ for the right figure (hence the right nulls are conservative, the left nulls are not). These figures show that (a) all considered methods control the FWER at level 0.2, (b) Online Sidak and Fallback provide negligible improvement compared to the naive Alpha-Spending, and (c) ADDIS-Spending is much more powerful than other algorithms in both settings.

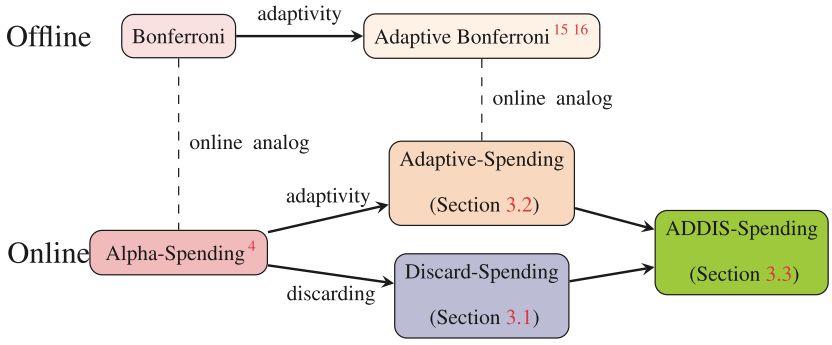


**Figure 2.** Historical context comparing offline and online FWER methods.

## 1.1 Paper outline

In Section 2, we extend some offline FWER methods to the online setting, and prove their FWER control. Then we derive ADDIS-Spending in Section 3, while we present its generalizations to handle local dependence in Section 3.4. Section 4 contains theoretical results about the statistical power of the online FWER control procedures (both known and newly derived), particularly Section 4.1 derives optimal choices for hyperparameters of naive Alpha-Spending within an idealized Gaussian setup, and Section 4.2 provides some theoretical justification of ADDIS-Spending being more powerful than naive Alpha-Spending. Numerical studies are included in Section 5 to empirically demonstrate the improved power of our new methods using both synthetic and real data. In the end, we also present several extensions of our main contribution, together with interesting follow-up directions for future work in Section 6.

# 2 Online extensions of classical offline FWER control methods

One may regard Alpha-Spending as an online generalization of Bonferroni correction. Specifically, for FWER level $\alpha$, given an infinite nonnegative sequence $\{\gamma_i\}_{i=1}^{\infty}$ that sums to one, Alpha-Spending tests individual hypothesis $H_i$ at level

$$\alpha_i := \alpha\gamma_i. \tag{2}$$

In this section, we seek online variants of other offline FWER methods such as Holm's[10] and Hochberg's methods[11]; fallback procedure[12] and alpha-recycling;[14] and Sidak correction.[13] Gladly, most of the methods (like Sidak, fallback and its generalization) can be trivially extended to online setting, while maintaining strong FWER control, while some of them (like Holm and Hochberg) which depend on ordering are hard to incorporate with the online setting. In the following, we present two general classes of algorithms which uniformly improve Bonferroni, together with numerical analysis of their performance. Specifically, those algorithms are Online Sidak (requires independence), an online analog of the Sidak correction[13]; and Online Fallback (works under arbitrary dependence), an online analog of the generalized fallback.[15] We expect more online procedures to be developed from offline methods in the future.

## 2.1 Online Sidak

Under independence, one well-known improvement of Bonferroni is the Sidak correction.[13] Given $m$ different null hypotheses and FWER level $\alpha$, the Sidak method uses the testing level $1-(1-\alpha)^{\frac{1}{m}}$ for each hypothesis. Analogously, Online Sidak chooses an infinite nonnegative sequence $\{\gamma_i\}_{i=1}^{\infty}$ which sums to one and tests each hypothesis $H_i$ at level

$$\alpha_i := 1-(1-\alpha)^{\gamma_i} \tag{3}$$

Just as Sidak is only slightly less stringent than Bonferroni, Online Sidak also improves very little of Alpha-Spending, as shown in Figure 3. In addition, Online Sidak also requires independence for valid FWER control, as stated below.

**Proposition 1.** *Online Sidak controls FWER in a strong sense if the null p-values are independent of each other, and is at least as powerful as the corresponding Alpha-Spending procedure.*

Proof. The probability of no false discovery among all infinite decisions is

$$\Pr\{V=0\} = \mathbb{E}\left[\prod_{i\in\mathcal{H}_0} 1_{P_i>\alpha_i}\right] \geq \prod_{i\in\mathcal{H}_0}(1-\alpha_i)$$
$$= \prod_{i\in\mathcal{H}_0}(1-\alpha)^{\gamma_i} = (1-\alpha)^{\sum_{i\in\mathcal{H}_0}\gamma_i} \geq 1-\alpha$$

Further, since $\alpha_i > \gamma_i\alpha$, Online Sidak is at least as powerful as the corresponding Alpha-Spending procedure.*2.2 Online fallback procedures*

The fallback procedure[12] and its graphical generalization[15] for offline FWER control can be easily extended to the online setting, improving the power of Alpha-Spending by "recycling" the significance level of previous rejections. Below, we briefly present the online variants of the generalized fallback,[15] specifically named Online Fallback.

### 2.2.1 Online fallback

For FWER level $\alpha$, Online Fallback chooses an infinite sequence $\{\gamma_i\}_{i=1}^{\infty}$ that sums to one, and tests each $H_i$ at level

$$\alpha_i := \gamma_i\alpha + \sum_{k=1}^{i-1} w_{k,i}R_k\alpha_k \tag{4}$$

where $R_i = 1_{P_i\leq\alpha_i}$, and $\{w_{ki}\}_{i=k+1}^{\infty}$ being some infinite sequence that is nonnegative and sums to one for all $k \in \mathbb{N}$. Specifically, we call $\{w_{ki}\}_{i=k+1}^{\infty}$ the transfer weights for the $k$-th hypothesis.

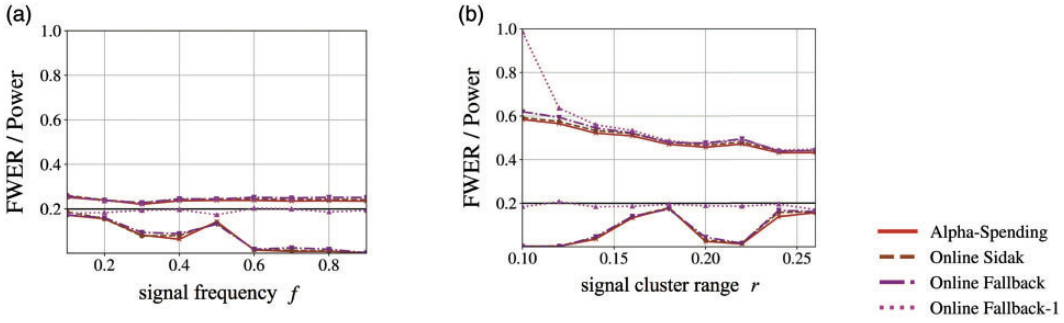


**Figure 3.** Statistical power and FWER for Alpha-Spending, Online Sidak, Online Fallback-1 and Online Fallback at target FWER level $\alpha = 0.2$ (solid black line). The curves above line 0.2 display the power of each method, while the lines below 0.2 display the FWER of each method. The specific experimental setting follows those in Section 5.1, with few alterations. The main experimental setting is described in Section 5.1, while we allow $\pi_A$ to be different for each hypotheses $H_i$, and denote it as $\pi_{Ai}$ instead. We set $\mu_N \equiv 0, \mu_A \equiv 4$ for all the figures, and we set $\pi_{A_i} = f$ for $i$ less than $\lfloor Tr \rfloor$ and $\pi_{A_i} = 0$ otherwise. Particularly, we set $f \in \{0.1, 0.2, \ldots, 0.9\}$ for the left figure, while $f \equiv 0.1$ for the right figure; and we set $r = 1$ for the first figure, and $r \in \{0.1, 0.12, \ldots, 0.26\}$ for the right figure. The underlying $\{\gamma_i\}_{i=1}^{\infty}$ is $\{6/\pi^2 i^2\}_{i=1}^{\infty}$ for all the methods, and particularly we set $\{w_{ki}\}_{i=k+1}^{\infty}$ with $w_{ki} = \gamma_{i-k}$ for all $k \in \mathbb{N}$ in Online Fallback. These figures show that (a) all the considered methods do control FWER at level 0.2; (b) the new methods (Online Sidak, Online Fallback-1 and Online Fallback) have almost the same power as Alpha-Spending given evenly distributed non-nulls; and (c) Online Fallback-1 and Online Fallback have noticeable power improvement over Alpha-Spending when non-nulls cluster in the beginning of the testing sequence, with greater improvement if the non-nulls are more compact in the beginning of the testing sequence.

Note that when we choose $\{w_{ki}\}_{i=k+1}^{\infty}$ with $w_{ki} = 1_{k=i-1}$, the individual testing level of Online Fallback reduces to

$$\alpha_i := \alpha\gamma_i + R_{i-1}\alpha_{i-1} \tag{5}$$

for all $i \geq 2$, and $\alpha_1 = \alpha\gamma_1$, which happens to be the offline fallback procedure[12] applying on the infinite hypotheses sequence. We specifically refer to this procedure as *Online Fallback-1*.

The following Proposition states the FWER control of Online Fallback and is proved in Appendix A. In fact, Online Fallback can be treated as an instantiation of a more general sequential rejection principle,[16] and Proposition 2 follows trivially from their results. However, we present a more tangible proof for Proposition 2 from a separate direction, since it allows us to get more insights about the connection between the offline methods and their online variants.

**Proposition 2.** *Online Fallback controls FWER for arbitrarily dependent p-values, and is at least as powerful as the corresponding Alpha-Spending procedure.*

Though Proposition 2 states the power superiority of Online Fallback over Alpha-Spending, one may have the intuition that the power improvement of Online Fallback over Alpha-Spending would be minor when one encounters randomly arriving signals, since much of the recycled significance levels are inevitably wasted on the nulls. Figure 3 demonstrates this intuition, where the left figure describes the setting of randomly arrived signals and plot against varied signal frequency $f$, while the right figure describes the setting of clustered signals at the beginning of the sequence and plot against varied cluster range $r$. As shown in the left figure of Figure 3, where the non-nulls are evenly distributed, Online Fallback has basically the same power with Alpha-Spending. Meanwhile, in the extreme case when the beginning of the sequence consists of only non-nulls (e.g. when $r = 1$ in the right figure), Online Fallback (especially Online Fallback-1) is much more powerful than Alpha-Spending, while this advantage diminishes fast as more non-nulls plug in the signal cluster (i.e. as $r$ increases).

The above new online FWER control methods are guaranteed to be uniformly more powerful than Alpha-Spending, though the improvements are usually minor, except for extreme cases when the non-nulls are clustered at the beginning of the testing sequence. This motivates the development of a new class of algorithms in the following section.

## 3 Adaptive discarding (ADDIS) algorithms

In Section 2, we refine Alpha-Spending by refining its offline variants. Since we find that the resulting methods rarely improve power much over Alpha-Spending, we refine Alpha-Spending from another angle in this section, directly addressing main sources of looseness in the proof of its FWER control. We end up with a series of new adaptive algorithms that are much more powerful than Alpha-Spending, though at the cost of requiring independence. We ease the requirement in Section 3.4 to allow the methods to handle a local (in time) dependence structure, which is a more reasonable dependence structure to assume for the online setting.[17]

First, we explain the looseness in the proof of FWER control of Alpha-Spending. Recall the Alpha-Spending procedure in equation (2). Its FWER is controlled because an even stronger error metric, the per-family wise error rate (PFER) is also controlled at $\alpha$

$$\text{PFER} := \mathbb{E}[V] = \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} 1_{P_i \leq \alpha_i}\right] = \sum_{i\in\mathcal{H}_0} \mathbb{E}[1_{P_i \leq \alpha_i}] \overset{(A)}{\leq} \sum_{i\in\mathcal{H}_0} \alpha_i \overset{(B)}{\leq} \sum_{i\in\mathbb{N}} \alpha_i = \alpha \tag{6}$$

(One can verify that FWER $\leq$ PFER by Markov's inequality.) The first obvious loose point in the proof above may be enforcing PFER control while aiming for FWER control. However, it turns out this looseness is negligible under independence, which we formally prove in Appendix I; in short, the improvement is of order $\alpha^2$, which is a lower order term, and this phenomenon can also be observed in Figure 3 where Online Sidak (FWER) has a very slight advantage over Alpha-Spending (PFER).

The two labeled inequalities in equation (6) introduce the main sources of looseness: the PFER of Alpha-Spending may be far less than $\alpha$ (and hence the power would be much lower than necessary) if either of the two inequalities is loose. Specifically, the first inequality (A) would be loose if the null $p$-values are very conservative, and the second inequality (B) would be loose if there are a large number of non-nulls. These facts expose two

weaknesses of Alpha-Spending: it will become suboptimal when it encounters conservative nulls or a constant fraction (like 20%) of non-nulls, and both are arguably quite likely in real applications. These weaknesses are also problems for the methods derived in Section 2, since they have almost the same power with Alpha-Spending in those simulations.

By addressing these two weaknesses, we develop a more powerful family of methods called ADDIS-Spending, which are adaptive discarding algorithms that not only benefit from adaptivity to the fraction of nulls, but also gain by exploiting conservative nulls (if they exist). Instead of directly presenting the ADDIS-Spending algorithm, we first introduce the idea of discarding and adaptivity in Section 3.1 and Section 3.2, respectively, each addressing one of the looseness mentioned above, and in Section 3.3 we combine the ideas together to develop our ADDIS-Spending algorithm, which addresses both loosenesses. There is a price to pay for these improvements. Alpha-Spending works even when the $p$-values are arbitrarily dependent, but the idea of discarding and adaptivity essentially requires independence between $p$-values. We ease this requirement later in Section 3.4, by generalizing our ADDIS-Spending algorithm to handle a local dependence structure.

Before we proceed, it is useful to set up some notation. Recall that $P_j$ is the $p$-value resulted from testing hypothesis $H_j$. Given infinite sequences $\{\tau_t\}_{t=1}^{\infty}$, $\{\lambda_t\}_{t=1}^{\infty}$ and $\{\alpha_t\}_{t=1}^{\infty}$, where every element is in (0, 1), define the indicators

$$S_j = 1_{P_j \le \tau_j}, \quad C_j = 1_{P_j \le \lambda_j}, \quad R_j = 1_{P_j \le \alpha_j} \tag{7}$$

which, respectively, indicate whether $H_j$ is selected for testing (used for adapting to conservative nulls), whether $H_j$ is a candidate for rejection (used for adapting to fraction of nulls), and whether $H_j$ is rejected. Accordingly define $R_{1:t} = \{R_1, \ldots, R_t\}$, $C_{1:t} = \{C_1, \ldots, C_t\}$ and $S_{1:t} = \{S_1, \ldots, S_t\}$. Similarly, let $\mathcal{R} = \{i \in \mathrm{N} : R_i = 1\}$, $\mathcal{C} = \{i \in \mathrm{N} : C_i = 1\}$, $\mathcal{S} = \{i \in \mathrm{N} : S_i = 1\}$. In what follows, we say $\alpha_t$, $\lambda_t$ and $\tau_t$ are "predictable with respect to some filtration $\mathcal{F}^t$", or just "predictable" for short, to mean that they are measurable with respect to $\mathcal{F}^{t-1}$, meaning that they are mappings from $\mathcal{F}^{t-1}$ to (0, 1). The form of $\mathcal{F}^{t-1}$ may change across algorithms and it is denoted as $\sigma(\cdot)$, that is a sigma field generated by certain random variables.

## 3.1 Discarding conservative nulls

Here, we develop a method named Discard-Spending to address the first looseness (A) in equation (6) due to overlooking the conservativeness of null $p$-values. Specifically, we consider the null $p$-values to be *uniformly conservative* as defined below, which include uniform nulls (the ideal case) and the majority of conservative nulls (realistic case).

**Definition 1.** *(Uniformly conservative) A null p-value P is said to be uniformly conservative if for all* $x, \tau \in [0, 1]$

$$\Pr\{P \le x\tau \,|\, P \le \tau\} \le x \tag{8}$$

The above condition is easy to interpret; for example when $\tau = 0.6$ and $x = 0.5$, it means the following—if we know that $P$ is at most 0.6, then it is more likely between $[0.3, 0.6]$ than in $[0, 0.3]$. As an obvious first example, uniform null $p$-values are uniformly conservative. As for more general examples, note that the definition in equation (8) is mathematically equivalent to requiring that for the CDF $F$ of the $p$-value $P$, we have

$$F(\tau x) \le xF(\tau) \quad \text{for all } x, \tau \in [0, 1] \tag{9}$$

Hence, null $p$-values with convex CDF are uniformly conservative.

Moreover, noting that monotonically nondecreasing density implies convex CDF for null $p$-values, Zhao et al.[3] presented the following Proposition 3.

**Proposition 3.** *For a one-dimensional exponential family with true parameter* $\theta \le \theta_0$*, the p-value resulting from the uniformly most powerful (UMP) test of* $H_0 : \theta \le \theta_0$ *versus* $H_1 : \theta > \theta_0$ *is uniformly conservative.*

In particular, the corresponding nulls in Proposition 3 are exactly uniform at the boundary of the null set ($\theta = \theta_0$), while conservative at the interior. Since the true underlying state of nature is rarely *exactly* at the boundary, it is common in practice to encounter uniformly conservative nulls that are indeed conservative.

One simple example following Proposition 3 is the one-sided test of Gaussian mean, where we test the null hypothesis $H_0 : \mu \leq 0$ against the alternative $H_1 : \mu > 0$ with observation $Z \sim N(\mu, 1)$. From Proposition 3, the $p$-value resulting from UMP test, that is $P = \Phi(-Z)$ is uniformly conservative, and is conservative if $\mu < 0$, while the conservativeness increases as $\mu$ decreases, where $\Phi$ is the standard Gaussian CDF. Another simple example is the two-sided test of Gaussian mean, where we test the null hypothesis $H_0 : |\mu| \leq \epsilon$ against the alternative $H_1 : |\mu| \geq \epsilon$ with observation $Z \sim N(\mu, 1)$. From Proposition 3, the $p$-value resulting from the UMP test, that is $P = 2\Phi(-|Z| + \epsilon)$, is uniformly conservative, and is conservative if $|\mu| < \epsilon$, while the conservativeness increases as $|\mu|$ decreases.

**Remark 1.** *The uniformly conservative condition is also known as the uniform conditional stochastic order (UCSO) relative to uniform distribution U(0, 1).*[3,18,19]

There have been some works addressing uniformly conservative $p$-values in the offline setting,[3,18] which both boil down to one simple idea: *discard* (do not test) large $p$-values, and test the others after some rescaling. In particular, Zhao et al.[3] used it in the global null test setting whileEllis et al.[18] used it with regard FWER/FDR control. It has recently been utilized for more powerful online FDR control by Tian and Ramdas.[9] Here we extend this *discarding* idea to online setting for FWER control, stated specifically as the following Discard-Spending algorithm.

#### 3.1.1 Discard-Spending

Recalling the definitions in and right after equation (7), we call any online FWER algorithm as a Discard-Spending algorithm if it updates the $\alpha_i$ in a way such that $\{\alpha_i\}_{i=1}^{\infty}$ satisfies the following conditions: (1) $\alpha_i$ is predictable, where the filtration $\mathcal{F}^{i-1} = \sigma(R_{1:i-1}, S_{1:i-1})$; (2) $\tau_i$ is predictable, and $\alpha_i < \tau_i$ for all $i \in \mathrm{N}$ and

$$\sum_{i \in \mathcal{S}} \frac{\alpha_i}{\tau_i} \equiv \sum_{i \in \mathrm{N}} \frac{\alpha_i}{\tau_i} S_i \leq \alpha \tag{10}$$

Note that condition (10) does not conflict with the predictability condition of $\{\tau_i\}_{i=1}^{\infty}$ and $\{\alpha_i\}_{i=1}^{\infty}$, even though it apparently requires knowing the whole sequence. In fact, since $\alpha_i, \tau_i, S_i \geq 0$, achieving condition (10) is equivalent to maintaining

$$\hat{\alpha}^{(t)} := \sum_{i=1}^{t} \frac{\alpha_i}{\tau_i} S_i \leq \alpha \tag{11}$$

for each $t \in \mathrm{N}$.

In other words, $\alpha_{t+1}, \tau_{t+1}$ are chosen such that $\hat{\alpha}^{(t)} + \alpha_{t+1} \leq \alpha$, depending only on information till time $t$. To better demonstrate how to construct such $\alpha_{t+1}, \tau_{t+1}$, we show an explicit example of Discard-Spending algorithm:

• *Choose a nonnegative sequence* $\{\gamma_i\}_{i=1}^{\infty}$ *that sums to one, given any predictable* $\{\tau_i\}_{i=1}^{\infty}$*, and we test every* $H_i$ *at the adapted level*

$$\alpha_i := \alpha \tau_i \gamma_{t(i)}, \quad \text{where } t(i) = 1 + \sum_{j<i} S_j. \tag{12}$$

It is easy to check that, by the above definition $\alpha_i < \tau_i$, and $\alpha_i$ is $\mathcal{F}^{i-1}$-measurable for all $i \in \mathrm{N}$. With simple algebra, we also have

$$\sum_{i \in \mathcal{S}} \frac{\alpha_i}{\tau_i} = \sum_{i \in \mathcal{S}} \alpha \gamma_{t(i)} = \sum_{i \in \mathrm{N}} \alpha \gamma_{t(i)} S_i = \alpha \sum_{t \in \mathrm{N}} \gamma_t = \alpha \tag{13}$$

where we use the observation that $t(i+1) = t(i) + 1$ if and only if $S_i = 1$ under our construction. Therefore, example (12) is a Discard-Spending algorithm.

On the other hand, condition (10) implies that Discard-Spending is equivalent to the following strategy: if $P_i > \tau_i$, then we do not test it (we *discard* it, which essentially means not to move forward in the $\{\alpha_i\}_{i=1}^{\infty}$ sequence), and if $P_i \le \tau_i$, we test $P_i$ at some level $\alpha_i \in (0,1)$ satisfying equation (10).

Next, we prove the claim that general Discard-Spending algorithms control FWER.

**Proposition 4.** *Discard-Spending controls PFER and thus FWER in a strong sense when the null p-values are uniformly conservative as defined in equation (8), while being independent of each other and of the non-nulls.*

Proof. We prove this theorem mainly using the law of iterated expectation and the uniform conservative property of null $p$-values. Recall that V is the number of false discoveries, which can be written as $V = \sum_{i \in \mathcal{H}_0} R_i S_i$, using the fact that $\alpha_i \le \alpha \le \tau_i$ by construction. Therefore

$$\begin{aligned} \mathbb{E}[V] &= \sum_{i \in \mathcal{H}_0} \mathbb{E}[R_i S_i] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[R_i S_i \mid S_i, \mathcal{F}^{i-1}\right]\right] \\ &= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[R_i \mid S_i = 1, \mathcal{F}^{i-1}\right] \Pr\{S_i = 1\}\right] \\ &= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\Pr\left\{\frac{P_i}{\tau_i} \le \frac{\alpha_i}{\tau_i} \mid P_i \le \tau_i, \mathcal{F}^{i-1}\right\} \Pr\{S_i = 1\}\right] \end{aligned} \tag{14}$$

where we used linearity of expectation twice.

Note that the property of uniformly conservative nulls in equation (8) gives us

$$\Pr\left\{\frac{P_i}{y} \le \frac{x}{y} \mid P_i \le y\right\} \le \frac{x}{y}, \text{ for all } i \in \mathcal{H}_0, \text{ and constants } x \le y \in [0,1] \tag{15}$$

Under the independent assumption among all $p$-values, the following variant of equation (15) also holds true

$$\Pr\left\{\frac{P_i}{y_i} \le \frac{x_i}{y_i} \mid P_i \le y_i, \mathcal{F}^{i-1}\right\} \le \frac{x_i}{y_i}, \text{ for all } i \in \mathcal{H}_0 \tag{16}$$

where $x_i$ and $y_i$ are some predictable random variables with respect to filtration $\mathcal{F}^i$ (i.e. $x_i, y_i \in \mathcal{F}^{i-1}$). Following a similar reasoning, the validity of null $p$-values can be rephrased under the independence assumption as

$$\Pr\left\{P_i \le z_i \mid \mathcal{F}^{i-1}\right\} \le z_i, \text{ for all } i \in \mathcal{H}_0, \text{ and nonnegative } z_i \in \mathcal{F}^{i-1} \tag{17}$$

Using the above observations, and the fact that $\alpha_i, \tau_i \in \mathcal{F}^{i-1}$, we have for $i \in \mathcal{H}_0$

$$\begin{aligned} \mathbb{E}\left[\Pr\left\{\frac{P_i}{\tau_i} \le \frac{\alpha_i}{\tau_i} \mid P_i \le \tau_i, \mathcal{F}^{i-1}\right\} \Pr\{S_i = 1\}\right] &\le \mathbb{E}\left[\frac{\alpha_i}{\tau_i} \Pr\{S_i = 1\}\right] \\ = \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_i}{\tau_i} 1_{P_i \le \tau_i} \mid S_i = 1, \mathcal{F}^{i-1}\right] \Pr\{S_i = 1\}\right] \\ = \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_i}{\tau_i} 1_{P_i \le \tau_i} \mid S_i, \mathcal{F}^{i-1}\right]\right] &= \mathbb{E}\left[\frac{\alpha_i}{\tau_i} 1_{P_i \le \tau_i}\right] \end{aligned} \tag{18}$$

where we use the law of iterated expectation in the last step. Plugging equation (18) back in equation (14), we have

$$\mathbb{E}[V] \le \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\alpha_i}{\tau_i} 1_{P_i \le \tau_i}\right] = \mathbb{E}\left[\sum_{i \in \mathcal{H}_0 \cap \mathcal{S}} \frac{\alpha_i}{\tau_i}\right] \overset{(i)}{\le} \mathbb{E}\left[\sum_{i \in \mathcal{S}} \frac{\alpha_i}{\tau_i}\right] \le \alpha \tag{19}$$

by construction. Therefore, PFER $\le \alpha$ and thus FWER $\le \alpha$ as claimed.

Discarding can lead to higher power when there are many conservative nulls. However, inequality (i) in equation (19) will be really loose if we have $|\mathcal{H}_0 \cap \mathcal{S}| \ll |\mathcal{S}|$, meaning that most of the contents of $\mathcal{S}$ are non-nulls, which is possible since the indices in $\mathcal{S}$ are those with small $p$-values. Figure 4 demonstrates both the strength and weakness of Discard-Spending, where we use $\alpha_i$ as described in the concrete example (12) mentioned above, and set $\tau_i = 0.5$ and $\gamma_i = \frac{6}{\pi^2 i^2}$ for all $i$.

## 3.2 Adaptivity to unknown proportion of nulls

Here, we develop a method to address the second looseness (B) in equation (6), which is due to lack of adjustment for the proportion of true nulls. Related ideas have been proposed during the development of offline FWER methods, which can be improved by incorporating an estimate of the true null proportion. This led to a series of adaptive methods like the adaptive Bonferroni,[20,21] and other generalizations like adaptive Sidak, adaptive Holm and adaptive Hochberg, which are rigorously proved to have FWER control.[22–24] Inspired by those efforts, we introduce the Adaptive-Spending procedure next, which could be regarded as an online variant of adaptive Bonferroni.

### 3.2.1 Adaptive-spending

Recalling the definitions in and right after equation (7), we call any online FWER algorithm as an Adaptive-Spending algorithm if it updates $\alpha_i$ in a way such that $\{\alpha_i\}_{i=1}^{\infty}$ satisfies the following conditions: (i) $\alpha_i$ is predictable, where the filtration $\mathcal{F}^{i-1} = \sigma(R_{1:i-1}, C_{1:i-1})$; (ii) for a predictable sequence $\{\lambda_i\}_{i=1}^{\infty}$, we have

$$\sum_{i \notin \mathcal{C}} \frac{\alpha_i}{1-\lambda_i} \equiv \sum_{i \in \mathrm{N}} \frac{\alpha_i}{1-\lambda_i}(1-C_i) \leq \alpha \tag{20}$$

Following the similar reasoning in Section 3.1, condition (20) does not conflict with the predictability condition of $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\alpha_i\}_{i=1}^{\infty}$, even though it apparently requires knowing the whole sequence. To better demonstrate how to construct such sequences in the online fashion, we show an explicit example of Adaptive-Spending algorithm:

• *Choose a nonnegative sequence* $\{\gamma_i\}_{i=1}^{\infty}$ *that sums to one, given any predictable* $\{\lambda_i\}_{i=1}^{\infty}$*, we test every* $H_i$ *at the adapted level*

$$\alpha_i := \alpha(1-\lambda_i)\gamma_{t(i)}, \quad \text{where } t(i) = 1 + \sum_{j<i}(1-C_j) \tag{21}$$

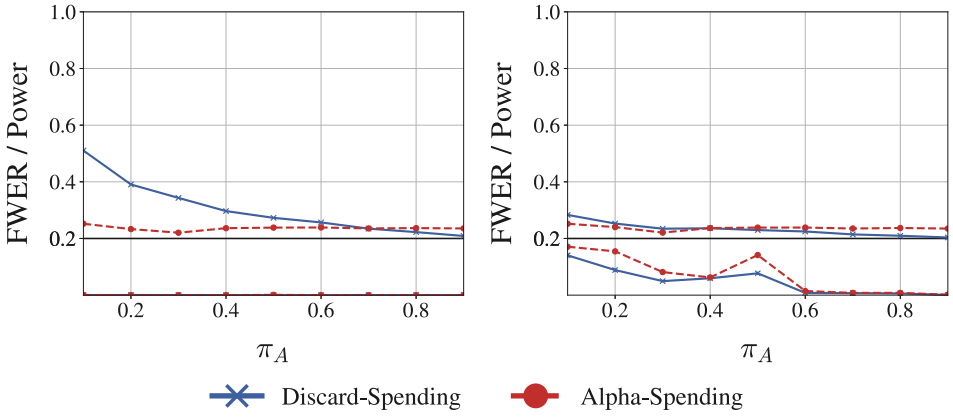


**Figure 4.** Statistical power and FWER versus fraction of non-null hypotheses $\pi_A$ for Discard-Spending and Alpha-Spending at target FWER level $\alpha = 0.2$ (solid black line). The curves above line 0.2 display the power of each method versus $\pi_A$, while the lines below 0.2 display the FWER of each method versus $\pi_A$. The experimental setting is described in Section 5.1: we set $\mu_A = 4$ for both figures, but $\mu_N = -2$ for the left figure and $\mu_N = 0$ for the right figure (hence the left nulls are conservative, the right nulls are not). These figures show that: (a) Discard-Spending do control FWER at level 0.2; (b) Discard-Spending is more powerful than naive Alpha-Spending when the nulls are conservative (as shown in the left figure); (c) Discard-Spending loses power when a high proportion of non-nulls is encountered (as shown in the right figure).

It is easy to check above that by the above construction, $\alpha_i$ is $\mathcal{F}^{i-1}$-measurable for all $i \in \mathrm{N}$. With simple algebra, we additionally have

$$\sum_{i \notin \mathcal{C}} \frac{\alpha_i}{1-\lambda_i} = \sum_{i \notin \mathcal{C}} \alpha \gamma_{t(i)} = \sum_{i \in \mathrm{N}} \alpha \gamma_{t(i)}(1-C_i) = \alpha \sum_{t \in \mathrm{N}} \gamma_t = \alpha \tag{22}$$

where we use the observation that $t(i+1) = t(i)+1$ if and only if $C_i = 0$ under our construction. Therefore, equation (21) is an Adaptive-Spending algorithm.

From condition (20) we can similarly simplify Adaptive-Spending as the following strategy: whenever $P_i \leq \lambda_i$, we don't lose any error budget for testing it at level $\alpha_i$; but whenever $P_i > \lambda_i$, we lose $\alpha_i/(1-\lambda_i)$ from our error budget.

Next, we prove the FWER control of Adaptive-Spending.

**Proposition 5.** *Adaptive-Spending controls PFER and thus FWER, when the null p-values are independent of each other and of the non-nulls.*

Proof. We prove the PFER (and hence FWER) control of Adaptive-Spending mainly using the law of iterated expectation and the fact that the null $p$-values are valid, i.e. $\Pr\{P_i \leq x\} \leq x$ for any $x \in [0,1]$ and $i \in \mathcal{H}_0$. Recalling that $V$ is the number of false discoveries, using the law of iterated expectation, we have

$$\mathbb{E}[V] = \mathbb{E}\left[\sum_{i \in \mathcal{H}_0} 1_{P_i \leq \alpha_i}\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[1_{P_i \leq \alpha_i} \mid \mathcal{F}^{i-1}\right]\right] \tag{23}$$

Note that the validity of null $p$-values gives us equation (17) as explained in the proof of Proposition 4, and therefore we have

$$\mathbb{E}\left[\mathbb{E}\left[1_{P_i \leq \alpha_i} \mid \mathcal{F}^{i-1}\right]\right] \overset{(i)}{\leq} \mathbb{E}[\alpha_i] \overset{(ii)}{\leq} \mathbb{E}\left[\alpha_i \mathbb{E}\left[\frac{1_{P_i > \lambda_i}}{1-\lambda_i} \mid \mathcal{F}^{i-1}\right]\right] = \mathbb{E}\left[\alpha_i \frac{1_{P_i > \lambda_i}}{1-\lambda_i}\right] \tag{24}$$

where the last equality follows from the law of iterated expectation. Plugging equation (24) back in equation (23), we have

$$\mathbb{E}[V] \leq \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\alpha_i \frac{1_{P_i > \lambda_i}}{1-\lambda_i}\right] \leq \mathbb{E}\left[\sum_{i \in \mathrm{N}} \alpha_i \frac{1_{P_i > \lambda_i}}{1-\lambda_i}\right] = \mathbb{E}\left[\sum_{i \notin \mathcal{C}} \frac{\alpha_i}{1-\lambda_i}\right] \leq \alpha \tag{25}$$

Therefore, PFER $\leq \alpha$ and thus FWER $\leq \alpha$ as claimed.

Adaptive procedures can improve power substantially if there is a non-negligible proportion of signals. However, their power can also suffer considerably if the null $p$-values are very conservative, since inequalities (i) and (ii) in equation (24) would be extremely loose when $\Pr\{P_i \leq x\} \ll x$ and $\Pr\{P_i > x\} \gg 1-x$. These strengths and weaknesses of Alpha-Spending are demonstrated in Figure 5, where we use $\alpha_i$ as described in the concrete example (21) mentioned above, and set $\lambda_i = 0.5$ and $\gamma_i = \frac{6}{\pi^2 i^2}$ for all $i \in \mathrm{N}$.

## 3.3 Combining the two ideas: ADDIS-Spending, an adaptive discarding alpha-spending algorithm

From the discussion in Sections 3.1 and 3.2, we find adaptivity and discarding both have their strength and weakness; however, these are complementary. Therefore, we next combine those two ideas for a more sophisticated online FWER control method. Specifically, we present the following ADDIS-Spending algorithm, where "ADDIS" stands for "ADaptive DIScarding". In the following Sections 4 and 5, we demonstrate the power superiority of ADDIS-Spending over Alpha-Spending, with both theoretical justifications in Section 4.2, and numerical analysis using synthetic and real data in Section 5.

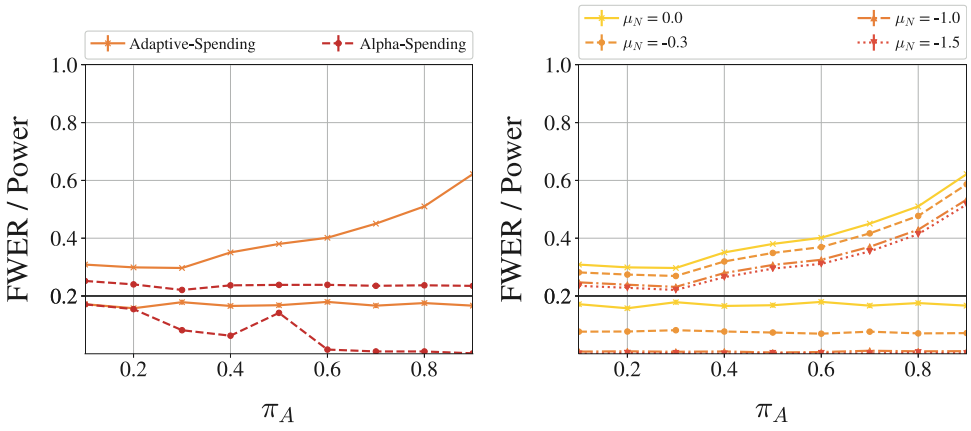


**Figure 5.** Statistical power and FWER versus fraction of non-null hypotheses $\pi_A$ for Adaptive-Spending, Alpha-Spending at target FWER level $\alpha = 0.2$ (solid black line). The curves above line 0.2 display the power of each method versus $\pi_A$, while the lines below 0.2 display the FWER of each methods versus $\pi_A$. The $p$-values are drawn in the same way as described in Section 5.1: we set $\mu_A = 4$ for both figures, but $\mu_N = 0$ for the left figure and $\mu_N \in \{0, -0.3, -1, -1, 5\}$ for the right figure (hence the left nulls are conservative, the right nulls are not). These figures show that: (a) Adaptive-Spending do control FWER at level 0.2; (b) Adaptive-Spending is more powerful than naive Alpha-Spending when there are many non-nulls (as shown in the left figure); (c) Adaptive-Spending loses power when many conservative nulls are encountered (as shown in the right figure).

### 3.3.1 ADDIS-Spending

Recall the definitions in and right after equation (7). We call any online FWER control method as an ADDIS-Spending algorithm if it updates individual testing level $\alpha_i$ in a way satisfying the following conditions: (1) $\alpha_i$ is predictable, where the filtration $\mathcal{F}^{i-1} = \sigma(R_{1:i-1}, C_{1:i-1}, S_{1:i-1})$; (2) predictable sequences $\{\tau_i\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$ are such that $\lambda_i < \tau_i$ and $\alpha_i < \tau_i$ for all $i$ and

$$\sum_{i \in \mathcal{S} \setminus \mathcal{C}} \frac{\alpha_i}{\tau_i - \lambda_i} \equiv \sum_{i \in \mathrm{N}} \frac{\alpha_i}{\tau_i - \lambda_i}(S_i - C_i) \leq \alpha \tag{26}$$

Following the similar reasoning in Section 3.1, condition (26) does not conflict with the predictability condition of $\{\tau_i\}_{i=1}^{\infty}$, $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\alpha_i\}_{i=1}^{\infty}$, even though it apparently requires knowing the whole sequence. To better demonstrate how to construct such sequences, we show an explicit example of ADDIS-Spending algorithm:

• *Choose a nonnegative sequence* $\{\gamma_i\}_{i=1}^{\infty}$ *that sums to one, given any predictable* $\{\tau_i\}_{i=1}^{\infty}$*, we test every* $H_i$ *at the adapted level*

$$\alpha_i := \alpha(\tau_i - \lambda_i)\gamma_{t(i)}, \quad \text{where } t(i) = 1 + \sum_{j<i}(S_j - C_j) \tag{27}$$

It is easy to check above that by the above construction $\alpha_i < \tau_i$, and $\alpha_i$ is $\mathcal{F}^{i-1}$-measurable for all $i \in \mathrm{N}$. With simple algebra, we additionally have

$$\sum_{i \in \mathcal{S} \setminus \mathcal{C}} \frac{\alpha_i}{\tau_i - \lambda_i} = \sum_{i \in \mathcal{S} \setminus \mathcal{C}} \alpha\gamma_{t(i)} = \sum_{i \in \mathrm{N}} \alpha\gamma_{t(i)}(S_i - C_i) = \alpha \sum_{t \in \mathrm{N}} \gamma_t = \alpha \tag{28}$$

where we use the observation that $t(i+1) = t(i) + 1$ if and only if $S_i = 1$ and $C_i = 0$ under construction. Therefore, equation (27) is an ADDIS-Spending algorithm.

ADDIS-Spending can be regarded as a unification of the Adaptive-Spending and Discard-Spending algorithms we mentioned earlier: when setting $\lambda_i \equiv 0$ for all $i$, ADDIS-Spending recovers Discard-Spending; and when setting $\tau_i \equiv 1$ for all $i$, ADDIS-Spending recovers Adaptive-Spending. At a high level, the advantages of ADDIS-Spending come from this unification of adaptivity and discarding. We obtained similar success in the ADDIS algorithm for online FDR control[9] suggesting that our work may be regarded as a variant of ADDIS for a more stringent error metric. Finally, we prove the FWER control of ADDIS-Spending.

**Theorem 1.** *ADDIS-Spending controls the PFER (hence FWER) when null p-values are uniformly conservative as defined in equation (8), while being independent of each other and of the non-nulls.*

Theorem 1 is actually a special case of a more general version Theorem 2 that we are going to introduce later in Section 3.4, where we relax the independence assumption to local dependence, which includes independence as a special case. Therefore, we omit the proof for the special case and present the proof for the more general version later.

## 3.4 Handling local dependence

In Section 3.3, we introduced ADDIS-Spending, a new powerful variants of Alpha-Spending, though at the cost of requiring independence among the $p$-values, while naive Alpha-Spending works even when $p$-values are arbitrarily dependent. Indeed, the assumption of independence is rarely met in real applications: tests that occur nearby in time may share the same dataset; null hypotheses are often constructed given the information of recent testing results, etc. On the other hand, arbitrary dependence between sequential $p$-values is also arguably unreasonable: the dataset used for testing or the testing results from the distant past is usually considered having no impact on the current testing. In light of this, we consider another dependence structure that is more realistic—local dependence, first proposed by Zrnic et al.,[17] and is defined as follows

$$\text{For all } i \in \mathrm{N}, \text{ there exists } L_i \in \mathrm{N}, \text{ such that } P_i \perp P_{i-L_i-1}, P_{i-L_i-2}, \ldots, P_1 \tag{29}$$

where $\{L_i\}_{i=1}^{\infty}$ is a sequence of constants that we refer to as lags. The lags can depend on the experiment being run and on the sources of data, but not on the data themselves. Implicitly, $P_i$ may be arbitrarily dependent on $P_{i-1}, \ldots, P_{i-L_i}$ and in particular, when $L_i \equiv 0$ for all i, assuming local dependence reduces to assuming independence. It may be simplest to think of $L_i \equiv L$ to be fixed, but formally we assume for simplicity that $L_{i+1} \leq L_i + 1$. For example, the latter requirement avoids the case that $P_5$ is independent of $P_1$, $P_2$ (if $L_5 = 2$) while $P_6$ is arbitrarily dependent on $P_1$, $P_2$ (if $L_6 = 5$). We refer readers to the paper[17] for more detailed definition and discussions.

Here, we give simple alterations of the procedures in Section 3 that allows them to deal with local dependence. The way we accomplish this is to follow the "principle of pessimism".[17] Specifically, this principle suggests ignoring what really happened in the previous $L_t$ steps when deciding what to do at time $t$, and hallucinate a pessimistic outcome for those steps instead. Formally, the alterations we made for procedures in Section 3 insist that

$$\alpha_i, \tau_i, \lambda_i \in \mathcal{F}^{i-L_i-1}, \text{ for all } i \in \mathrm{N} \tag{30}$$

while still satisfying the other requirements in the corresponding original definitions.

As for concrete examples to implement the altered procedures described above, we present the altered concrete example of ADDIS-Spending in equation (27): we choose $\{\gamma_i\}_{i=1}^{\infty}$ as an infinite nonnegative sequence that sums to one, we test each $H_i$ at predictable level

$$\alpha_i := \alpha(\tau_i - \lambda_i)\gamma_{t(i)}, \quad \text{where } t(i) = 1 + L_i \wedge (i-1) + \sum_{j < i - L_i} S_j - C_j \tag{31}$$

Note that when $L_i = 0$ for all $i$, that is the local dependence structure reduces to independence, the above modified procedures reduce to ADDIS-Spending in equation (27).

We now present the PFER (and hence FWER) control of altered ADDIS-Spending for local dependence in Theorem 2, which is proved in Appendix B.

**Theorem 2.** *Altered ADDIS-Spending controls PFER (and hence FWER) in a strong sense when the null p-values are uniformly conservative as defined in equation* (8) *and follow the local dependence defined in equation* (29).

## 4 Statistical power

We now study the statistical power of Alpha-Spending and ADDIS-Spending under an idealized Gaussian setting with randomly arriving signals. Specifically, in Section 4.1 we examine the power of Alpha-Spending, and we derive some optimal choices of the underlying $\{\gamma_i\}_{i=1}^{\infty}$ in a certain range, for either fixed or varying signal strength and density. Then we provide theoretical justification for the benefits of adaptivity and discarding in Section 4.2: we prove that if $\{\gamma_i\}_{i=1}^{\infty}$ lies in the aforementioned optimal range, ADDIS-Spending is more powerful than Alpha-Spending.

Before we proceed with the analysis of power, it is useful to set up a few definitions.

**Definition 2.** (q-series and log-q-series) For any $q > 1$, we call an infinite sequence $\{\gamma_i\}_{i=1}^{\infty}$ which is nonnegative and sums to one as q-series if $\gamma_i \propto i^{-q}$ for all i, and similarly, as a log-q-series if $\gamma_i \propto 1/i\log^q i$ for all i.

**Definition 3.** (Gaussian mean testing problem) We call the problem of testing a possibly infinite sequence of hypotheses $\{H_i\}_{i=1}^{\infty}$ as Gaussian mean testing problem, if each observation $Z_i$ follows the following mixed distribution

$$Z_i = \begin{cases} X_i + \mu_A & \text{with probability } \pi_A; \\ X_i + \mu_N, & \text{with probability } 1 - \pi_A \end{cases}$$

where constants $\mu_A > 0$, $\mu_N \leq 0$, $\pi_A \in (0,1)$ for all $i$, $X_i \overset{iid}{\sim} N(0,1)$, and we test the null hypothesis $H_i : \mu_i \leq 0$, where $\mu_i = \mathbb{E}[Z_i]$.

Recalling Section 3.1, if the $p$-values calculated are one-sided that is $P_i = \Phi^{-1}(-Z_i)$, then we know that the nulls are uniformly conservative as defined in equation (8), strictly conservative when $P_i = \Phi^{-1}(-Z_i)$, and uniform when $\mu_N = 0$.

In the following, we only consider the online Gaussian mean testing problem described above. We compare the algorithms that are presented as concrete examples of each method, which are formulas (2) and (27) for Alpha-Spending and ADDIS-Spending, respectively, with the same underlying sequence $\{\gamma_i\}_{i=1}^{\infty}$. Also, for simplicity, we use the number of true discoveries $D$ as one of the performance measures. Note that $D$ is the numerator inside the expectation of the power function (1). Since the denominator inside the expectation of the power function in equation (1) remains the same for different algorithms given the same testing sequence, the expectation of numerator $D$ may arguably serve as a nice substitution for power function in respect of comparison. Hence we refer $\mathbb{E}[D]$ also as the power of online FWER control methods in this section.

### 4.1 Getting optimal power using naive Alpha-Spending

In this section, we derive optimal choices of $\{\gamma_i\}_{i=1}^{\infty}$ in the range of $q$-series for Alpha-Spending with regard to the Gaussian mean testing problem in Definition 3. As we discussed before, the expectation of number of true discoveries serves as a reasonable measurement for comparing the power of testing procedures. Recall that $D$ is the number of true discoveries. In the Gaussian mean testing problem, we have

$$\begin{aligned} \mathbb{E}[D] &= \mathbb{E}\left[\sum_{i=1}^{\infty} 1_{P_i \leq \alpha_i, i \in \mathcal{H}_0^c}\right] =^{(i)} \sum_{i=1}^{\infty} \mathbb{E}\left[1_{P_i \leq \alpha_i, i \in \mathcal{H}_0^c}\right] = \sum_{i=1}^{\infty} \Pr\left\{P_i \leq \alpha_i, i \in \mathcal{H}_0^c\right\} \\ &= \sum_{i=1}^{\infty} \pi_A \Phi(\Phi^{-1}(\alpha_i) + \mu_A) = \sum_{i=1}^{\infty} \pi_A \Phi(\Phi^{-1}(\alpha\gamma_i) + \mu_A) \end{aligned} \tag{32}$$

where $\Phi$ is the standard Gaussian CDF, and (i) is true due to the fact that each entry in the summation is nonnegative, and the last step uses the fact that $\alpha_i = \alpha\gamma_i$ for all $i$ in Alpha-Spending. Additionally, for each $N \in \mathbb{N}$, we denote $\mathbb{E}_N[D]$ as the expectation of true discoveries among the first $N$ hypotheses, which means

$$\mathbb{E}_N[D] := \sum_{i=1}^{N} \pi_A \Phi(\Phi^{-1}(\alpha\gamma_i) + \mu_A) \tag{33}$$

It is obvious that the power of Alpha-Spending does not depend on $\mu_N$, therefore we only consider how to choose $\{\gamma_i\}_{i=1}^{\infty}$ to optimize the power given different $\mu_A$ and $\pi_A$. Specifically, we derive that for the Gaussian mean testing problem in Definition 3 using Alpha-Spending, the optimal sequence $\{\gamma_i\}_{i=1}^{\infty}$ in the range of $q$-series will be $q = 1^+$ for any choice of $\mu_A > 0$ and $\pi_A \in (0,1)$. This result provides some heuristic of choosing $\{\gamma_i\}_{i=1}^{\infty}$: one should resort to log-$q$-series for higher power when applicable. The formal results are stated in the following Theorem 3.

**Theorem 3.** *Recall the definition of $\mathbb{E}[D]$ and $\mathbb{E}_N[D]$ in equations (32) and (33). For Alpha-Spending (2) at level $\alpha < 1/2$, if the underlying sequence $\{\gamma_i\}_{i=1}^{\infty}$ is a $q$-series where $q > 1$, then for the Gaussian mean testing problem in Definition 3, we have*

*(a) For $N \geq 2$, $\mathbb{E}_N[D]$ is a unimodal function, first increasing with $q$ and then decreasing with $q$. Additionally, defining*

$$q^*(N, \mu_A) \equiv \text{argsup}_{q>1} \mathbb{E}_N[D] \tag{34}$$

*we have that $q^*(N, \mu_A)$ is monotonically decreasing with $N$ for any $\mu_A > 0$, and*

$$\lim_{N\to\infty} q^*(N, \mu_A) = 1 \tag{35}$$

*• $\mathbb{E}[D]$ is finite for any fixed $q$, and is a function monotonically decreasing with $q$.*

Theorem 3 in fact suggests that the slower $\{\gamma_i\}_{i=1}^{\infty}$ is decaying, the higher the power will get, which corresponds to our intuition that the power will be higher if we protect the ability to detect the signals in the long run, or we try not to run out of our total budget too fast. However, if the testing process stops at a certain point, then sequence $\{\gamma_i\}_{i=1}^{\infty}$ that decays too slow will hurt the power, since the testing will not get the benefit from the long run. This trade-off implies that there is an optimal sequence, which is not decaying too slow or too fast, making the testing process achieve the highest power. Theorem 3 provides theoretical verification for those intuitions and is proved in Appendix D.

Therefore, when we have no prior information on the hypotheses, which means we could only treat the non-null fraction and the non-null mean as some fixed arbitrary value, we could always resort to the sequence $\{\gamma_i\}_{i=1}^{\infty}$ that sums to one with slower decay rate to obtain higher power of Alpha-Spending. For example, among $q$-series, we should choose $q$ as close to one as possible, and one should resort to log-$q$-series for higher power.

The above results are all in the regime of fixed signal strength and density, which is a bit unrealistic in practice, though potentially suitable for deriving interpretable heuristics. For completeness, we also go beyond the setting of fixed signal strength and density: in Appendix E, we consider different $\mu_{Ai}$ and $\pi_{Ai}$ for each $i$ in Definition 3, and we show that, if $\{\mu_{Ai}\}_{i=1}^{\infty}$ and $\{\pi_{Ai}\}_{i=1}^{\infty}$ satisfy some reasonable conditions, then there exists a function $h$ with closed form, such that $\gamma_i = h(\pi_{Ai}, \mu_{Ai})$ achieves the highest power. We refer readers to Appendix E for details.

## 4.2 The adaptive discarding methods are more powerful

In Section 4.1, we showed that the optimal choice of $\{\gamma_i\}_{i=1}^{\infty}$ for Alpha-Spending in the range of $q$-series and log-$q$-series for the Gaussian mean testing problem with fixed signal strength and density lies in the regime of log-$q$-series. Here, we show that in this regime, ADDIS-Spending is provably more powerful than Alpha-Spending.

For simplicity, we consider fixed $\tau_i$ and $\lambda_i$, that is $\tau_i = \tau$ and $\lambda_i = \lambda$ for all $i \in \mathrm{N}$, and we denote the number of discoveries from ADDIS-Spending as $D_{\text{ADDIS}}(\lambda, \tau)$, and $D_{\text{spend}}$ for Alpha-Spending. Below, we demonstrate that as long as the hyper-parameters $\lambda$ and $\tau$ are reasonably chosen (i.e. in ranges we derived in the following, which depend only on the distribution of $p$-values), ADDIS-Spending is guaranteed to be more powerful than Alpha-Spending.

**Theorem 4.** *For the Gaussian mean testing problem in Definition 3, if the underlying $\{\gamma_i\}_{i=1}^{\infty}$ is a log-$q$-series as defined in Definition 2 with $q > 1$, then there exists some $c^*$ such that with probability one,*

$$\mathbb{E}[D_{\text{ADDIS}}(\lambda, \tau)] \geq \mathbb{E}[D_{\text{spend}}] \quad \text{for all } \lambda \in [0, c^*) \text{ and all } \tau \in (c^*, 1], \text{s.t.} \tau - \lambda < 1 \tag{36}$$

*Additionally, $c^*$ increases with $\pi_A$, $\mu_A$ and $\mu_N$, and equals one when $\mu_N = 0$.*

The corresponding proof is in Appendix F. Theorem 4 indicates that, as long as we have reasonable prior information about signal strength, density and conservativeness of nulls, we can utilize them for better design of

hyperparameters $\lambda$ and $\tau$ in ADDIS-Spending, such that higher power over Alpha-Spending can be achieved. As for a more tangible recommendation for users, we observe the combination $\tau = 0.8$ and $\lambda = \alpha\tau$ perform well in our empirical studies as we discuss next, and is the most robust choice across various signal settings as shown in the greedy evaluation in Appendix J.2. Appendix J.2 can be taken as a reference table for scientists to make the best use of ADDIS-Spending in practice, as we cover settings of various signal density and nulls conservativeness there. As Theorem 4 provides the theoretical justification for the benefits of discarding and adaptivity in terms of power, in the following Section 5 we provide numerical analysis with both simulations and real data example to confirm these benefits.

# 5 Numerical studies

## 5.1 Simulations

In this section, we provide some numerical experiments to compare the performance of ADDIS-Spending, Discard-Spending, Adaptive-Spending, and Alpha-Spending. In particular, for each method, we provide empirical evaluations of its power while ensuring that the FWER remains below a chosen value.

Specifically, in the following, we aim to control the FWER under $\alpha = 0.2$ and estimate the FWER and power by averaging over 2000 independent trials. The constant sequences $\tau_i \equiv 0.8$ in Discard-Spending, $\lambda_i \equiv \alpha$ in Adaptive-Spending, and $\tau_i \equiv 0.8, \lambda_i \equiv \alpha\tau = 0.16$, in ADDIS-Spending for all $i \in \mathbb{N}$ were found to be generally successful, so as our default choice in this section and we drop the index for simplicity. Additionally, we choose the infinite sequence $\gamma_i \propto 1/(i+1)\log(i+1)^2$ for all $i \in \mathbb{N}$ as default, which could be substituted by any constant infinite sequence that is nonnegative and sums to one.

In what follows, we show the power superiority of ADDIS-Spending over all other three methods, especially under settings with both nonnegligible number of signals and conservative nulls. Specifically, we consider the simple experimental setup of Gaussian mean testing problem in Definition 3 with $T = 1000$ components, where the nulls are uniformly conservative from the discussion in Section 3.1.

We ran simulations for $\mu_N \in \{0, -0.5, -1, -1.5\}$, $\mu_A \in \{4, 5\}$, and $\pi_A \in \{0.1, 0.2, \ldots, 0.9\}$, to see how the changes in conservativeness of nulls and true signal fraction may affect the performance of algorithms. The results are shown in Figure 6, which indicates that (1) FWER is under control for all methods in all settings; (2) ADDIS-Spending enjoys appreciable power increase as compared to all the other three methods in all settings; (3) the more conservative the nulls are (the more negative $\mu_N$ is), or the higher the fraction of non-nulls is, the more significant the power increase of ADDIS-Spending is.

Here for conciseness, we only present results with some default parameter choices under canonical settings, which turn out working pretty well in establishing the strength of our methods and confirm the theoretical results in Section 4.2, though they may not be optimal in obtaining high power. We conduct more thorough empirical studies with different parameters choices under other practical cases in Appendix J, where we demonstrate consistent power superiority of ADDIS-Spending.

In particular, to provide more intuition of suitable hyper-parameters, we show the influence of different $\{\gamma_i\}_{i=1}^{\infty}$ choices empirically in Appendix J.1, which confirms our theoretical results in Section 4.1 that slow-decaying sequences lead to higher power; we explore greedily on feasible $\tau$ and $\lambda$ choices in Appendix J.2 to depict optimal combination of them, and find that promising region corresponds to our theoretical finding in Section 4.2. On the other hand, in order to show generality of our methods under other practical settings, in Appendix J.3 we demonstrate consistent success of our methods in the two-sided problem, a common problem formulation in many areas like genomics; we show the consistent validity and superiority of our methods under more strict FWER level requirement in Appendix J.4, a desired requirement in areas that need serious risk control; we reassure similar success of our methods under various strength of local dependence in Appendix J.5, a common dependence structure that happens in practice when independence assumption breaks.

## 5.2 Application to the international mouse phenotyping consortium

In the following, we introduce the application of ADDIS-Spending to real-life data about International Mouse Phenotyping Consortium (IMPC). The IMPC coordinates a large study to functionally annotate every protein coding gene by exploring the impact of the gene knockout on the resulting phenotype. Statistically speaking, for each phenotype $i$, people test the null hypothesis $H_j^{(i)}$: "the knockout of gene $j$ will not change the phenotype $i$" versus its alternative, via comparing the unmutated mouse (control case) to the mouse with gene $j$ knockout. Since

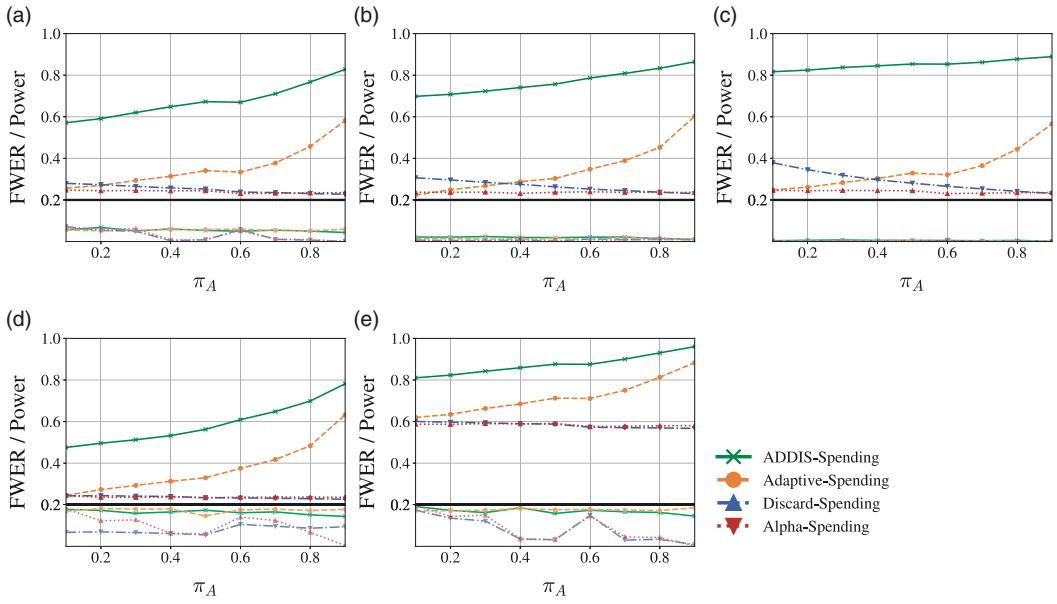


**Figure 6.** Statistical power and FWER versus fraction of non-null hypotheses $\pi_A$ for ADDIS-Spending, Discard-Spending, Adaptive-Spending and Alpha-Spending at target FWER level $\alpha = 0.2$ (solid black line). The lines above the solid black line are the power of each method versus $\pi_A$, and the lines below are the FWER of each method versus $\pi_A$. The $p$-values are drawn using the Gaussian model as described in the text, while we set $\mu_N = -0.5$ in plot (a), $\mu_N = -1$ in plot (b), $\mu_N = -1.5$ in plot (c), and $\mu_N = 0$ in plots (d) and (e); and we set $\mu_A = 4$ in plots (a) to (d), and $\mu_A = 5$ in plot (e). Therefore nulls in (a), (b) and (c) are conservative, and the conservativeness is increasing, while the nulls in (d) and (e) are not conservative (uniform).

the dataset and resulting hypotheses constantly grow as new knockouts are studied, and a positive test outcome could lead to some following up medical research that is hard to be revised afterwards, it is natural to view this as an online testing problem as the ones considered in the previous sections.

We follow the analysis done by Karp et al.,[25] which resulted in a set of $p$-values for testing genotype effects. These are available at the Zenodo repository,[b] organized by Robertson.[26] This particular dataset admits a natural local dependence structure: the hypotheses are tested in small batches, with each batch using a different group of mice. Figure 7(a) demonstrates this local dependence structure via showing the first 5000 $-\log_{10}$ transformed $p$-values: the transformed $p$-values are ordered by the time that its corresponding data samples are collected, and the adjacent batches are distinguished using different colors.

Due to this online nature and local dependence structure of the data, we hence apply the modified version of ADDIS-Spending described in Section 3.4, using lags $L_i$ that corresponds to the size of blocks that $p$-value $P_i$ belongs to. Since we do not know the underlying truth, we only report the number of discoveries and argue that the corresponding FWER are all under control following our theoretical results. Figure 7(b) shows the power advantage of ADDIS-Spending over Alpha-Spending and Online-Fallback, where we use the same underlying $\{\gamma_i\}_{i=1}^{\infty}$ sequence with $\gamma_i \propto 1/i^{1.1}$ (similar qualitative behavior is shown in Appendix J.1 with $\gamma_i \propto 1/i^2$), and the default setting $\tau = 0.8, \lambda = 0.16$ for ADDIS-Spending in comparison.

The above real data example again supports our key idea: utilizing the independence or local dependence structure to incorporate with adaptability and discarding can improve the power of online testing procedure much.

## 6 Conclusion and discussion

Modern biology studies often require testing hypotheses in a sequential manner, and how to control familywise error rate in this setting leads to the statistical problem of online FWER control. This paper derives new algorithms for online FWER control, a problem for which no systematic treatment exists in the literature to the best of our knowledge. While we describe several new methods, each improving on Alpha-Spending (online Bonferroni) in different ways, the most promising of these in experiments seems to be ADDIS-spending, a new adaptive discarding algorithm that adapts to both unknown number of non-nulls and conservativeness of the nulls.

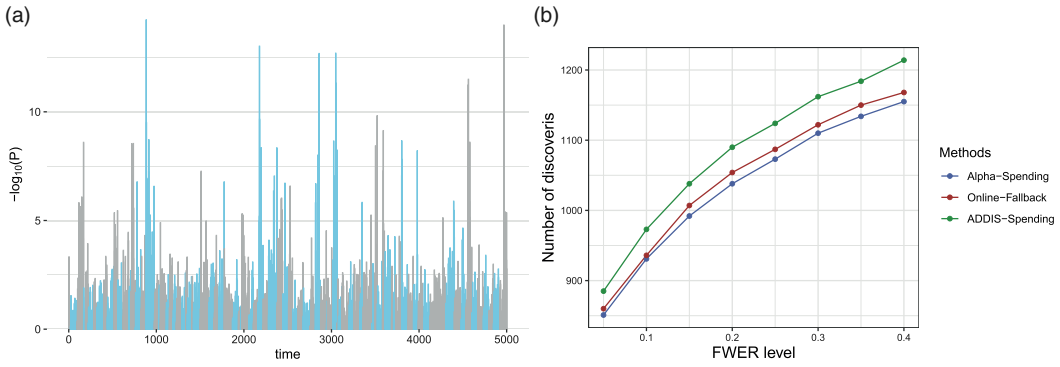


**Figure 7.** Figure (a) shows the barplot of $-\log_{10}$ transformed $p$-values, where the $p$-values are ordered in time, and each adjacent batche is distinguished with different colors. Figure (b) shows the number of discoveries versus FWER level, using different algorithms (Alpha-Spending, Online-Fallback, ADDIS-Spending).

Though we find that ADDIS-Spending is the most promising method within current practices, we are also wondering whether there exists universal refinements over ADDIS-Spending. It is known that for any offline global null testing methods, there exists a closure such that the power of it is unimprovable. So we are wondering whether there exists similar logic for online multiple testing. We provide some initial attempts on developing the variant of the closure principle for online multiple testing in Appendix H. Several questions still remain: Is our proposed principle essentially unimprovable? Also, is the closure of an online method still be an online method? If not always, then what are the cases in which it is? We leave these as open questions for future work.

The application of adaptivity and discarding go much beyond the main contribution of this paper: they can also be applied to methods in Section 2 to develop more powerful variants. We provide some concrete examples and corresponding proofs for their FWER control in Appendix G for interested readers.

Throughout this paper, we mainly discuss the online algorithms for controlling FWER. In real applications, many prefer to control $k$-FWER instead, in order to obtain a less stringent error control. The $k$-FWER is defined as $\Pr\{V \geq k\}$, which reduces to FWER as $k = 1$. It is straightforward that for any methods that have PFER control, changing the sum of the test levels to $k\alpha$ will assure $k$-FWER controlled at level $\alpha$, simply using Markov's equality. Therefore, all our new algorithms that provably have PFER control may easily be extended to $k$-FWER control methods.

## 7 Code and data availability

The code to reproduce all figures is accessible at Github repository, and the real dataset is available at Zenodo repository, organized by Robertson.[26] Additionally, an R package called onlineFDR[27] developed by David Robertson and the authors of this paper (among others), contains current state of arts in all aspect of online multiple testing, including online FWER control and also the new algorithms proposed here.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: AR would like to acknowledge NSF CAREER Award 1945266.

### Supplementary material

Supplementary material for this paper is available online.

### ORCID iD

Jinjin Tian https://orcid.org/0000-0003-3537-6430

## Acknowledgements

We thank Jelle Goeman and David S. Robertson for a careful reading and useful suggestions, particularly pointing the authors to related work,[16,18] and the available IMPC dataset.[26]

## Notes

a. The marginal distribution of a null $p$-value $P$ is typically assumed to satisfy $\Pr\{P \leq x\} \leq x$ for all $x \in [0,1]$. Ideally, equality holds, but in practice, null $p$-values are often *conservative*, meaning that $\Pr\{P \leq x\} \ll x$.
b. The Zenodo repository can be accessed at https://zenodo.org/record/2396572.

## References

1. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B (Stat Methodol)* 1995; **57**: 289–300.
2. Storey J. A direct approach to false discovery rates. *J Royal Stat Soc Ser B (Stat Methodol)* 2002; **64**: 479–498.
3. Zhao Q, Small DS and Su W. Multiple testing when many $p$-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J Am Stat Assoc* 2018; **114**: 1291–1304.
4. Foster D and Stine R. $\alpha$-investing: a procedure for sequential control of expected false discoveries. *J Royal Stat Soc Ser B (Stat Methodol)* 2008; **70**: 429–444.
5. Aharoni E and Rosset S. Generalized $\alpha$-investing: definitions, optimality results and application to public databases. *J Royal Stat Soc Ser B (Stat Methodol)* 2014; **76**: 771–794.
6. Javanmard A and Montanari A. Online rules for control of false discovery rate and false discovery exceedance. *Ann Stat* 2018; **46**: 526–554.
7. Ramdas A, Yang F, Wainwright M, et al. Online control of the false discovery rate with decaying memory. *In: Advances in neural information processing systems*, Long Beach, United States, 4-9 December 2017, pp.5655–5664.
8. Ramdas A, Zrnic T, Wainwright M, et al. SAFFRON: an adaptive algorithm for online control of the false discovery rate. In: *Proceedings of the 35th international conference on machine learning*, Stockholm, Sweden, 10-15 July 2018, vol. 80, pp.4286–4294.
9. Tian J and Ramdas A. ADDIS: adaptive algorithms for online FDR control with conservative nulls. In: *Proceedings of the 33rd conference on neural information processing systems*, Vancouver, Canada, 9-12 December 2019.
10. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; **6**: 65–70.
11. Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**: 800–802.
12. Wiens BL and Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *J Biopharmaceut Stat* 2005; **15**: 929–942.
13. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* 1967; **62**: 626–633.
14. Burman CF, Sonesson C and Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Stat Med* 2009; **28**: 739–761.
15. Bretz F, Maurer W, Brannath W, et al. A graphical approach to sequentially rejective multiple test procedures. *Stat Med* 2009; **28**: 586–604.
16. Goeman JJ and Solari A. The sequential rejection principle of familywise error control. *Ann Stat* 2010; **38**: 3782–3810.
17. Zrnic T, Ramdas A and Jordan M. Asynchronous online testing of multiple hypotheses. *arxiv preprint arxiv 1812.05068* 2018.
18. Ellis JL, Pecanka J and Goeman J. Gaining power in multiple testing of interval hypotheses via conditionalization. *Biostatistics* 2020; 21: e65–e79.
19. Whitt W. Uniform conditional stochastic order. *J Appl Probabil* 1980; **17**: 112–123.
20. Schweder T and Spjøtvoll E. Plots of $p$-values to evaluate many tests simultaneously. *Biometrika* 1982; **69**: 493–502.
21. Hochberg Y and Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med* 1990; **9**: 811–818.
22. Finner H and Gontscharuk V. Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J Royal Stat Soc: Ser B (Stat Methodol)* 2009; **71**: 1031–1048.
23. Guo W. A note on adaptive Bonferroni and Holm procedures under dependence. *Biometrika* 2009; **96**: 1012–1018.
24. Sarkar SK, Guo W and Finner H. On adaptive procedures controlling the familywise error rate. *J Stat Plan Inference* 2012; **142**: 65–78.
25. Karp NA, Mason J, Beaudet AL, et al. Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat Commun* 2017; **8**: 1–12.
26. Robertson DS, Wildenhain J, Javanmard A, et al. onlineFDR: an R package to control the false discovery rate for growing data repositories. *Bioinformatics* 2019; **35**: 4196–4199.
27. Robertson DS, Ramdas A, Javanmard A, et al. *onlineFDR: Online FDR control*, 2019. R package 1.3.6, https://dsrobertson.github.io/onlineFDR/index.html (accessed February 2020).