The Annals of Statistics
2020, Vol. 48, No. 6, 3465–3487
https://doi.org/10.1214/19-AOS1938

SIMULTANEOUS HIGH-PROBABILITY BOUNDS ON THE FALSE DISCOVERY PROPORTION IN STRUCTURED, REGRESSION AND ONLINE SETTINGS

BY EUGENE KATSEVICH[1] AND AADITYA RAMDAS[2]

[1]*Department of Statistics and Data Science, Carnegie Mellon University, ekatsevi@andrew.cmu.edu*
[2]*Department of Statistics and Data Science, Carnegie Mellon University, aramdas@cmu.edu*

While traditional multiple testing procedures prohibit adaptive analysis choices made by users, Goeman and Solari (*Statist. Sci.* **26** (2011) 584–597) proposed a simultaneous inference framework that allows users such flexibility while preserving high-probability bounds on the false discovery proportion (FDP) of the chosen set. In this paper, we propose a new class of such simultaneous FDP bounds, tailored for nested sequences of rejection sets. While most existing simultaneous FDP bounds are based on closed testing using global null tests based on sorted p-values, we additionally consider the setting where side information can be leveraged to boost power, the variable selection setting where knockoff statistics can be used to order variables, and the online setting where decisions about rejections must be made as data arrives. Our finite-sample, closed form bounds are based on repurposing the FDP estimates from false discovery rate (FDR) controlling procedures designed for each of the above settings. These results establish a novel connection between the parallel literatures of simultaneous FDP bounds and FDR control methods, and use proof techniques employing martingales and filtrations that are new to both these literatures. We demonstrate the utility of our results by augmenting a recent knockoffs analysis of the UK Biobank dataset.

1. Introduction.

1.1. *Multiple testing and exploration.* Consider testing a set of hypotheses $\mathcal{H} = \{H_1, \dots, H_n\}$, which we identify with $[n] \equiv \{1, \dots, n\}$. The false discovery proportion (FDP) of a rejection set $\mathcal{R} \subseteq [n]$ is defined

$$\mathrm{FDP}(\mathcal{R}) \equiv \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}|} \equiv \frac{V}{R},$$

where $\mathcal{H}_0 \subseteq \mathcal{H}$ is the set of nulls and $\mathrm{FDP}(\mathcal{R}) \equiv 0$ when $\mathcal{R} = \varnothing$ by convention (we use the $\equiv$ symbol for definitions). Based on the data at hand, a multiple testing procedure returns a rejection set $\mathcal{R}^*$, and the Type-I error of such a procedure can be evaluated via different properties of the FDP distribution. For example, the false discovery rate (FDR) is defined as the mean of the FDP (Benjamini and Hochberg (1995)) and the false discovery exceedance (FDX) is defined as the probability that FDP exceeds a pre-chosen threshold γ (Lehmann and Romano (2005)):

$$\mathrm{FDR} \equiv \mathbb{E}\big[\mathrm{FDP}(\mathcal{R}^*)\big] \quad \text{and} \quad \mathrm{FDX} \equiv \Pr\big\{\mathrm{FDP}(\mathcal{R}^*) > \gamma\big\}.$$

A multiple testing procedure is said to control an error rate if it falls below a prespecified level; for example, a procedure controls the FDR at level q if $\mathrm{FDR} \le q$.

Received March 2018; revised December 2019.
MSC2020 subject classifications. Primary 62J15, 62F03; secondary 60G42, 62P10.
Key words and phrases. False discovery rate, FDR, multiple testing, post hoc confidence bounds, false discovery exceedance, FDX, uniform martingale concentration.

This multiple testing paradigm has been very successful as the workhorse of scientific discovery for the past several decades. However, Goeman and Solari (2011) (GS) argued that this prevailing paradigm may not be flexible enough to accommodate for the exploratory nature of modern large-scale data analysis: target levels for the FDP like q or γ are chosen in advance, and rejection sets obtained from FDR- or FDX-controlling procedures cannot be grown or shrunk without invalidating their guarantees. Therefore, the paradigm leaves little room for scientists seeking to use their domain expertise to adaptively select a rejection set while maintaining valid inferential guarantees. They may attempt to do so by applying a multiple testing procedure with different nominal levels and choosing one of the resulting rejection sets post hoc. They may also exclude rejected hypotheses that do not align with their scientific priors, or include unrejected hypotheses that were close to the rejection threshold. However, these practices may lead to an excess of false positives.

Motivated by these considerations, GS proposed a complementary *simultaneous inference* paradigm. In this paradigm, one constructs FDP upper bounds $\overline{\mathrm{FDP}}(\mathcal{R})$ that hold uniformly across all sets $\mathcal{R}$ with high probability:

$$\Pr\bigl\{\mathrm{FDP}(\mathcal{R}) \le \overline{\mathrm{FDP}}(\mathcal{R}) \text{ for all } \mathcal{R} \subseteq \mathcal{H}\bigr\} \ge 1-\alpha. \tag{1}$$

Such bounds allow the scientist to inspect any pairs $(\mathcal{R}, \overline{\mathrm{FDP}}(\mathcal{R}))$ and freely choose the rejection set $\mathcal{R}^*$ whose content and FDP bound suits them. Given the simultaneous nature of statement (1), the upper bound on FDP continues to hold on the chosen set despite the user's data-dependent decision:

$$\begin{aligned}&\Pr\bigl\{\mathrm{FDP}(\mathcal{R}^*) \le \overline{\mathrm{FDP}}(\mathcal{R}^*)\bigr\}\\ &\quad\ge \Pr\bigl\{\mathrm{FDP}(\mathcal{R}) \le \overline{\mathrm{FDP}}(\mathcal{R}) \text{ for all } \mathcal{R} \subseteq \mathcal{H}\bigr\} \ge 1-\alpha.\end{aligned}$$

GS obtain such bounds by building on the closed testing principle, where a *local test* $\phi_{\mathcal{R}}$ (i.e., a test of the global null for a restricted set of hypotheses) is performed for each subset of hypotheses $\mathcal{R} \subseteq \mathcal{H}$. The results of all these local tests are aggregated to form a bound $\overline{\mathrm{FDP}}$ that provably satisfies (1). Since then, there has been much exciting work on new algorithms and computational shortcuts for simultaneous FDP control, mostly based on closed testing, but these have been somewhat disconnected from the parallel growth in the FDR literature.

1.2. *A new class of simultaneous FDP bounds.* In this paper, we show that a variety of FDR procedures can be repurposed to obtain simultaneous FDP bounds, establishing a novel connection between FDR control and simultaneous FDP control. In particular, note that many FDR algorithms implicitly construct a path, or nested sequence of n potential rejection sets

$$\Pi \equiv (\mathcal{R}_0, \dots, \mathcal{R}_n), \quad \text{such that} \quad \varnothing \equiv \mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_n \subseteq [n].$$

Then an estimate of the FDP

$$\widehat{\mathrm{FDP}}(\mathcal{R}_k) \equiv \frac{a_0 + \widehat{V}(\mathcal{R}_k)}{|\mathcal{R}_k|} \tag{2}$$

is constructed for each $\mathcal{R}_k \in \Pi$, where $\widehat{V}(\mathcal{R}_k)$ is an estimate of $V(\mathcal{R}_k) \equiv |\mathcal{R}_k \cap \mathcal{H}_0|$ and $a_0 \ge 0$ is an additive regularization constant. This estimate is then used to obtain a cutoff point

$$k^* \equiv \max\bigl\{k : \widehat{\mathrm{FDP}}(\mathcal{R}_k) \le q\bigr\}, \tag{3}$$

based on which the rejection set $\mathcal{R}^* \equiv \mathcal{R}_{k^*}$ is defined.

Repurposing the path Π and the estimate $\widehat{V}$, we propose the bound

$$\overline{\mathrm{FDP}}(\mathcal{R}_k) \equiv \frac{\overline{V}(\mathcal{R}_k)}{|\mathcal{R}_k|}; \quad \overline{V}(\mathcal{R}_k) = \bigl\lfloor c(\alpha) \cdot \bigl(a + \widehat{V}(\mathcal{R}_k)\bigr)\bigr\rfloor, \tag{4}$$

where $c(\alpha)$ are tight, explicit, dimension-independent constants such that

$$\Pr\{\mathrm{FDP}(\mathcal{R}) \le \overline{\mathrm{FDP}}(\mathcal{R}) \text{ for all } \mathcal{R} \in \Pi\} \ge 1-\alpha, \tag{5}$$

as long as the p-values satisfy an independence assumption. The constant $c(\alpha)$ depends implicitly on the regularization $a > 0$, which does not need to be the same as the original regularization a_0. Usually, we set $a = 1$.

If desired, $\overline{\mathrm{FDP}}$ can also be extended to all sets $\mathcal{R} \subseteq \mathcal{H}$ to obtain a bound of the form (1) through the process of *interpolation* (Blanchard, Neuvial and Roquain (2020), Goeman, Hemerik and Solari (2019)), where logical relationships among hypotheses are leveraged to combine a set of simultaneous FDP bounds. In fact, interpolation can actually be used to improve our bounds on the path as well, since some bounds on the path may be tighter than others. In particular, we may freely replace $\overline{V}$ with

$$\begin{aligned}\overline{V}^{\mathrm{interp}}(\mathcal{R}_k) &\equiv \min_j\{|\mathcal{R}_k \setminus \mathcal{R}_j| + \overline{V}(\mathcal{R}_j)\} \\ &= \min\Big(|\mathcal{R}_k| - \max_{j\le k}\{|\mathcal{R}_j| - \overline{V}(\mathcal{R}_j)\}, \min_{j\ge k}\overline{V}(\mathcal{R}_j)\Big).\end{aligned} \tag{6}$$

Blanchard, Neuvial and Roquain (2020) proposed the first expression above to interpolate FDP bounds for nested rejection sets (see their Proposition 2.5), and the second expression shows that we may compute these interpolated bounds in linear time by maintaining a cumulative maximum of lower bounds on the number of true positives and a cumulative minimum of upper bounds on the number of false positives.

Figure 1 summarizes the proposed FDP bounds and how they compare and contrast to FDR methods and GS's simultaneous inference paradigm based on closed testing.

We prove our bounds by developing a simple yet versatile proof technique—based on a martingale argument rather different from those commonly used in the FDR literature—to

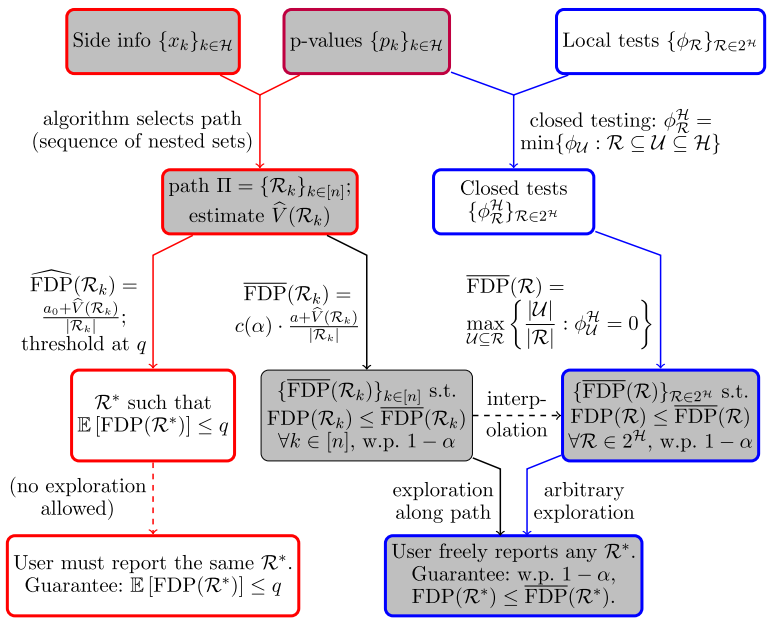

FIG. 1. *Schematic for proposed FDP bounds (shaded gray nodes), in the context of the usual FDR control framework (nodes with red borders) and the GS closed testing framework for simultaneous FDP control (nodes with blue borders). The proposed bounds borrow the path construction from FDR procedures to leverage side information, while obtaining simultaneous guarantees like the GS approach to permit exploration.*

obtain tight nonasymptotic bounds for the probability that the stochastic process $|\mathcal{R}_k \cap \mathcal{H}_0|$ of false discoveries hits certain boundaries. This technique is inspired by the proof of FDR control for the multilayer knockoff filter (Katsevich and Sabatti (2019)).

Several simultaneous bounds $\overline{\text{FDP}}$ for the sets $\mathcal{R}_t \equiv \{j : p_j \leq t\}$ are already available (Blanchard, Neuvial and Roquain (2020), Genovese and Wasserman (2006), Meinshausen (2006), Meinshausen and Rice (2006), Hemerik, Solari and Goeman (2019)), in addition to bounds valid for all subsets (van der Laan, Dudoit and Pollard (2004), Goeman and Solari (2011)). However, all of these bounds treat p-values exchangeably, whereas the link we establish between FDR and simultaneous FDP control allows us to leverage the rich recent literature (e.g., Lei and Fithian (2018), Li and Barber (2017)) on incorporating side information into FDR procedures to obtain powerful simultaneous FDP bounds. Importantly, our bounds apply also to the knockoffs procedure (Barber and Candès (2015), Candès et al. (2018)) for high-dimensional variable selection, which produces an ordered set of independent "one-bit p-values" to which we may apply one of our bounds. We note that a line of work has considered a form of simultaneous inference in the context of post-selection inference for linear models (Bachoc, Preinerstorfer and Steinberger (2016), Berk et al. (2013), Kuchibhotla et al. (2020)).

They also apply to the online setting, where p-values come in a stream and decisions to accept or reject must be made before seeing future data. Our results can also be used as diagnostic tools for FDR procedures: one can run a FDR procedure at a certain level and then obtain a valid $1 - \alpha$ confidence bound on the FDP of the resulting rejection set. Finally, all of our bounds (4) have an appealingly simple closed form.

Next, we preview our simultaneous FDP bound for the knockoffs procedure and demonstrate its utility on a large genome-wide association study data set (Section 2). Then we state our main results and provide a high level proof sketch in Section 3. In Section 4, we compare and contrast our theoretical results with those in the FDR literature. We then compare the performance of our simultaneous FDP bounds with existing alternatives via numerical simulations (Section 5) and then conclude the paper in Section 6. The code to reproduce our numerical simulations and data analysis is available online at https://github.com/ekatsevi/simultaneous-fdp.

2. An illustration with real data. Before formally stating and proving our bounds, we first illustrate their utility in the context of an application to genome-wide association studies (GWAS). The goal of GWAS is to identify the genetic factors behind various human traits. For this purpose, genotype and trait data are collected from large cohorts of individuals and then scanned for association. The recently compiled UK Biobank resource (Bycroft et al. (2018)) has data on a half a million individuals.

GWAS represents a vast variable selection problem, with genotypes viewed as covariates and the trait as the outcome. Since nearby genotypes are strongly correlated with each other, the units of inference usually are spatially localized genomic regions instead of individual genetic variants (i.e., variables are grouped before testing). The *knockoffs* framework (Barber and Candès (2015)) for variable selection with FDR control has been proposed to analyze GWAS data (Sesia, Sabatti and Candès (2019)) and has recently been applied to several phenotypes in the UK Biobank (Sesia et al. (2019)).

The knockoffs procedure falls into the class of FDR procedures introduced in the previous section. A set of *knockoff statistics* $W_1, \ldots, W_p$ are constructed for each group of genetic variants, with the property that the distribution of W_k is symmetric about the origin for null groups k. On the other hand, knockoff statistics for nonnull groups should be large and positive. Therefore, an ordering $\pi(1), \ldots, \pi(p)$ of the groups is constructed by sorting the knockoff statistics by decreasing magnitude. The kth rejection set $\mathcal{R}_k$ along the path is then

defined as the set of groups among the first k in the ordering whose knockoff statistics have positive signs:

$$\mathcal{R}_k \equiv \{\pi(j) \le k : \operatorname{sign}(W_{\pi(j)}) > 0\}.$$

FDR control is proved for regularization $a_0 = 1$ and the estimate

$$\widehat{V}(\mathcal{R}_k) \equiv |\{\pi(j) \le k : \operatorname{sign}(W_{\pi(j)}) < 0\}|, \tag{7}$$

which leverages the sign-symmetry of W_k for null k.

Our theoretical result (Corollary 1 in Section 3) shows that the bound

$$\overline{V}_{\text{knockoff}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\alpha^{-1})}{\log(2-\alpha)} \cdot (1 + |\{\pi(j) \le k : \operatorname{sign}(W_{\pi(j)}) < 0\}|) \right\rfloor, \tag{8}$$

that is, expression (4) with $a = a_0 = 1$, $c(\alpha) = \frac{\log(\alpha^{-1})}{\log(2-\alpha)}$, and $\widehat{V}$ as in definition (7), satisfies the uniform coverage statement (5). For $\alpha = 0.05$, we have $c(\alpha) \approx 4.5$. In other words, *inflating the knockoffs FDP estimate by 4.5 allows us to upgrade from bounding the FDP of one set on average to confidently bounding the true FDP across the entire path*.

To illustrate the utility of this result, we apply it to the analysis of the platelet count trait in the UK Biobank data, borrowing the knockoff statistics that were made publicly available by Sesia et al. (2019) at https://msesia.github.io/knockoffzoom/ukbiobank.html. While several correlation cutoffs were used to create groups in Sesia et al. (2019), here we consider the lowest resolution groups (of average width 0.226 megabases), whose size corresponds roughly to that yielded by current GWAS methodologies. In Figure 2, we plot $\overline{\text{FDP}}$ (obtained from interpolating the bounds (8) via (6)) and $\widehat{\text{FDP}}$ as a function of the rejection set size. The dashed line shows the FDR target level $q = 0.1$ used by Sesia et al. (2019). The $\widehat{\text{FDP}}$ curve crosses this threshold at k^* with $|\mathcal{R}_{k^*}| = 1460$. By comparison, the $\overline{\text{FDP}}$ curve is (necessarily) more conservative, but clearly yields informative FDP bounds for many rejection sets. It crosses the line $q = 0.1$ at $|\mathcal{R}_{k^*}| = 814$, meaning that we are 95% confident that at least 90% of the top 814 genomic loci are associated with platelet count.

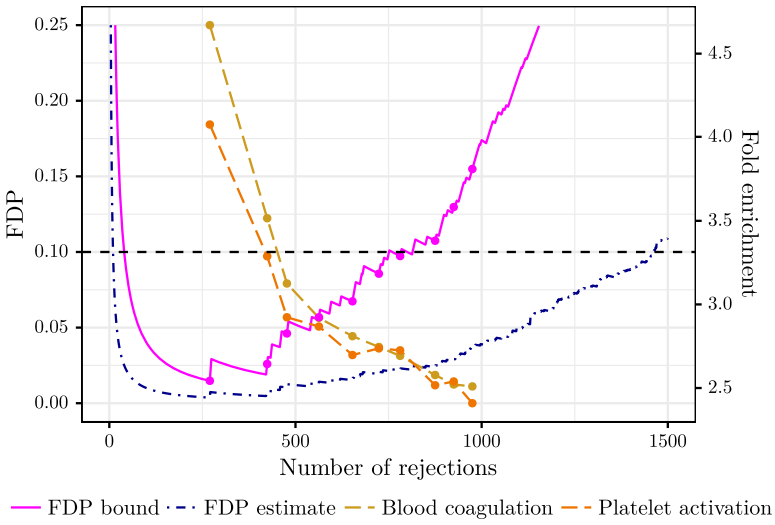

FIG. 2. *Knockoffs FDP estimate (dark blue) and proposed FDP upper bound (magenta) for GWAS analysis of platelet count, with highlighted points indicating interesting rejection sets. The degree of over-representation of two relevant Gene Ontology terms (orange and goldenrod) among genes in the neighborhood of genomic regions defined by each interesting rejection set.*

TABLE 1
Number of knockoffs discoveries for different traits in the UK Biobank, based on original analysis targeting FDR ≤ 0.1 *and proposed bounds with three FDP thresholds*

		With probability 0.95		
Trait	FDR < 0.1	FDP < 0.2	FDP < 0.1	FDP < 0.05
height	3284	2178	1672	1097
body mass index	1804	1162	755	498
platelet count	1460	1076	814	540
systolic blood pressure	722	463	327	208
cardiovascular disease	514	296	102	0
hypothyroidism	212	140	83	0
respiratory disease	176	111	83	0
diabetes	50	45	0	0

Importantly, though, we can do much more than this. Instead of committing to $q = 0.1$ before seeing the data, we can explore several rejection sets along the knockoffs path, examining their content and FDP bound. One strategy we might take is to choose a set of points along the knockoffs path that represent different compromises between FDP bound and rejection set size; these points are highlighted along the $\overline{\text{FDP}}$ curve in Figure 2. For example, the leftmost highlighted point represents a rejection set of size 270 with an FDP bound of 0.015 (with 95% confidence).

A domain expert might inspect each of these points and choose one that makes the most sense. In genetics, a common first step to evaluate a set of discoveries is to see whether they fit with known associations. The Gene Ontology, or GO (Ashburner et al. (2000)) is a collection of biological processes, each annotated with a set of genes known to be involved in that process. Given a set of genomic regions, the GREAT (McLean et al. (2010)) tool computes the "enrichment" (i.e., overrepresentation) of genes annotated to any given GO term falling in those genomic regions. For the platelet count trait, we would expect associated regions to be overrepresented for genes annotated to processes like "blood coagulation" or "platelet activation." We computed the fold enrichment (degree of overrepresentation) for these two terms, shown in Figure 2 as dashed goldenrod and orange lines. The fold enrichment generally decreases as we increase the size of the rejection set, corresponding to our intuition that the strongest signals are generally in regions previously known to be associated with platelet count. By juxtaposing domain-specific annotations with simultaneous FDP bounds, a plot like Figure 2 would already go a long way towards helping a domain expert decide on a biologically meaningful rejection set with a statistically sound Type-I error guarantee.

Finally, to quantify the price we pay for this extra flexibility, we consider several traits analyzed by Sesia et al. (2019) and compare the numbers of rejections we get for the original analysis (FDR ≤ 0.1) with the numbers we get for controlling FDP at various levels based on $\overline{\text{FDP}}_{\text{knockoff}}$. The results are shown in Table 1. As we can see, there is certainly a trade-off between analytical flexibility and statistical power. However, at least in this dataset, we can still make substantial numbers of discoveries while enjoying the benefit of improved flexibility.

Having previewed the utility of our theoretical results on a real dataset, in the next section we formally state these results.

3. Main results. In this section, we present a set of paths Π along with corresponding bounds $\overline{\text{FDP}}$ of the form (4) and state conditions under which the guarantee (5) holds. In fact, for brevity we will specify only the numerator $\overline{V}$ of $\overline{\text{FDP}}$. As discussed in the Introduction,

both Π and $\widehat{V}$ will be borrowed directly from existing FDR procedures. We provide bounds for both the *batch* and *online* settings. In the batch setting, there is a finite number of hypotheses $H_1, \ldots, H_n$ for which the p-values are available all at once; in the online setting, where there is an infinite stream of hypotheses, which arrive one at a time and a decision must be made about each hypothesis as soon as its p-value arrives. The proofs for all our results are provided in the Supplementary Material (Katsevich and Ramdas (2020)), but a sketch of the main idea is given in Section 3.3. Finally, recall that all the following bounds can be improved in linear time via interpolation (equation (6)), albeit losing their interpretable closed form.

3.1. *FDP bounds in the batch setting.* Here, we have a fixed, finite set of hypotheses $H_1, \ldots, H_n$ and a set of p-values $p_1, \ldots, p_n$. To construct a path, consider first ordering the hypotheses in some way $\pi(1), \pi(2), \ldots, \pi(n)$, constructing π to encourage nonnulls to appear near the beginning of the order. Then, define a p-value cutoff $p_* \in (0, 1]$. We form a path Π by traversing the ordering and choosing hypotheses whose p-values passed the cutoff:

$$\Pi \equiv (\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_n) \quad \text{such that} \quad \mathcal{R}_k \equiv \{\pi(j) : j \le k,\, p_{\pi(j)} \le p_*\}. \tag{9}$$

There are three ways of defining the path Π:

1. *sort*: π is formed by sorting p-values; in this case usually $p_* \equiv 1$.
2. *preorder*: π is fixed ahead of time using prior knowledge.
3. *interact*: π is built on the fly using prior knowledge and p-values.

Next, we elaborate on these path constructions in the batch setting and present FDP bounds for each of them.

3.1.1. *Sorted path.* Ordering hypotheses by p-value, $p_{\pi(k)} = p_{(k)}$, and setting $p_* = 1$ leads to

$$\mathcal{R}_k = \{j : p_j \le p_{(k)}\}. \tag{10}$$

This is the most common path construction among multiple testing procedures, serving as the basis for the Benjamini–Hochberg (BH) algorithm and many other step up/down algorithms (e.g., Benjamini and Liu (1999), Gavrilov, Benjamini and Sarkar (2009)). It is the obvious choice when no side information is available. Storey, Taylor and Siegmund (2004) formulated the BH algorithm in terms of the FDR control paradigm described in Section 1.2, with $\widehat{V}_{\text{sort}}(\mathcal{R}_k) \equiv n \cdot p_{(k)}$ and $a_0 = 0$:

$$\widehat{\text{FDP}}_{\text{sort}}(\mathcal{R}_k) \equiv \frac{n \cdot p_{(k)}}{|\mathcal{R}_k|}. \tag{11}$$

The following theorem presents our FDP bounds (4) for this path, based on the estimate $\widehat{V}_{\text{sort}}$.

THEOREM 1. *Let $\mathcal{R}_k$ be defined via (10), and let*

$$\overline{V}_{\text{sort}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{\log(1 + \log(\frac{1}{\alpha}))} \cdot (1 + n \cdot p_{(k)}) \right\rfloor.$$

If the null p-values are independent and stochastically larger than uniform, that is, $\Pr\{p_j \le s\} \le s$ for all $j \in \mathcal{H}_0$ and $s \in [0, 1]$, then the uniform bound (5) holds for all $\alpha \in (0, 0.31]$, that is,

$$\Pr\{\text{FDP}(\mathcal{R}_k) \le \overline{\text{FDP}}_{\text{sort}}(\mathcal{R}_k) \text{ for all } k \in [n]\} \ge 1 - \alpha.$$

REMARK 1. In Theorem 1, we require that $\alpha \leq 0.31$. However, strong numerical evidence shows that the bound is valid for all α. The restriction on α is an artifact of our proof and does not represent an intrinsic breaking point of the bound. Despite this limitation in our proof, the range $\alpha \leq 0.31$ includes most confidence levels that would be used in practice (although the case $\alpha = 0.5$ might be of interest to bound the median of the FDP distribution and $\alpha = 1$ of interest to bound the null proportion).

3.1.2. *Preordered path.* The *preordered setting* applies when prior information (e.g., data from a similar experiment) sheds light on which hypotheses are more likely to be nonnull, so a good ordering π is known in advance. Several FDR methodologies taking advantage of pre-specified orderings have been developed; G'Sell et al. (2016) and Li and Barber (2017) build paths using $p_* = 1$ while Barber and Candès (2015) and Lei and Fithian (2016) use $p_* \in (0, 1)$.

For the case $p_* = 1$, we use a construction from the accumulation test of Li and Barber (2017): an *accumulation function* h is a function $h : [0, 1] \to \mathbb{R}_+$ that is nondecreasing and integrates to 1. Then we define

$$\widehat{V}_{\text{preorder-acc}}(\mathcal{R}_k) \equiv \sum_{j=1}^{k} h(p_{\pi(j)}). \tag{12}$$

Alternatively, for $p^* \in (0, 1)$, we can follow Selective SeqStep (Barber and Candès (2015)) and Adaptive SeqStep (Lei and Fithian (2016)) to define

$$\widehat{V}_{\text{preorder-sel}}(\mathcal{R}_k) \equiv \sum_{j=1}^{k} \frac{p_*}{1-\lambda} I(p_{\pi(j)} > \lambda), \tag{13}$$

where $\lambda \geq p_*$. The following theorem presents our FDP bounds (4) for the preordered setting for the cases $p_* = 1$ and $p_* \in (0, 1)$, which rely on estimates (12) and (13), respectively.

THEOREM 2. *Fix $a > 0$ and assume the null p-values are independent and stochastically larger than uniform. Given a prior ordering π, let*

$$\mathcal{R}_k \equiv \{\pi(j) : j \leq k, p_{\pi(j)} \leq p_*\}.$$

1. *Set $p_* = 1$, choose a (possibly unbounded) accumulation function h, and define*

$$\overline{V}^h_{\text{preorder-acc}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{a \log((\int_0^1 \alpha^{h(u)/a}\, du)^{-1})} \cdot \left(a + \sum_{j=1}^{k} h(p_{\pi(j)})\right) \right\rfloor.$$

Then the uniform bound (5) holds for all $\alpha \in (0, 1)$:

$$\Pr\{\text{FDP}(\mathcal{R}_k) \leq \overline{\text{FDP}}^h_{\text{preorder-acc}}(\mathcal{R}_k) \text{ for all } k \in [n]\} \geq 1 - \alpha.$$

Moreover, if $\sup_{u\in[0,1]} h(u) \equiv B < \infty$, then we may instead use

$$\overline{V}^B_{\text{preorder-acc}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{a \log((1 - \frac{1-\alpha^{B/a}}{B})^{-1})} \cdot \left(a + \sum_{j=1}^{k} h(p_{\pi(j)})\right) \right\rfloor.$$

2. *Set $p_* \in (0, 1)$ and fix $\lambda \geq p_*$. Define*

$$\overline{V}^B_{\text{preorder-sel}} \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{a \log(1 + \frac{1-\alpha^{B/a}}{B})} \cdot \left(a + \sum_{j=1}^{k} \frac{p_*}{1-\lambda} I(p_{\pi(j)} > \lambda)\right) \right\rfloor,$$

where $B \equiv \frac{p_*}{1-\lambda}$. *Then uniform bound* (5) *holds for all* $\alpha \in (0, 1)$:

$$\Pr\{\mathrm{FDP}(\mathcal{R}_k) \le \overline{\mathrm{FDP}}^{B}_{\text{preorder-sel}}(\mathcal{R}_k) \text{ for all } k \in [n]\} \ge 1 - \alpha.$$

As we previewed in Section 2, we can apply this bound to the knockoff filter (Barber and Candès (2015)), a variable selection methodology based on the idea of creating a *knock-off variable* for each original variable, and then using these knock-offs as controls for the originals. Instead of p-values, the knock-off filter produces *knock-off statistics* W_j for each variable j. These are constructed so that

$$\{\mathrm{sign}(W_j)\}_{j\in\mathcal{H}_0} \perp\!\!\!\perp \{|W_j|\}_{j\in[p]}, \{\mathrm{sign}(W_j)\}_{j\notin\mathcal{H}_0}; \qquad \{\mathrm{sign}(W_j)\}_{j\in\mathcal{H}_0} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2). \tag{14}$$

The signs of the knock-off statistics are therefore a set of independent "one-bit p-values," to which the above theorem applies.

COROLLARY 1. *Let* $W_1, \ldots, W_p$ *be a set of knock-off statistics satisfying property* (14). *Let* π *be the ordering corresponding to sorting* W_j *by decreasing magnitude, and define* $\mathcal{R}_k = \{\pi(j) \le k : \mathrm{sign}(W_{\pi(j)}) > 0\}$. *Then bound* (5) *holds for*

$$\overline{V}_{\text{knockoff}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\alpha^{-1})}{a \log(2 - \alpha^{1/a})} \cdot \left(a + \left|\{\pi(j) \le k : \mathrm{sign}(W_{\pi(j)}) < 0\}\right|\right) \right\rfloor.$$

PROOF. Define $p_j = 1/2$ for $W_j < 0$ and $p_j = 1$ for $W_j > 0$. By property (14), it is easy to see that these p-values are independent of the ordering π (so the ordering can be treated as fixed) and satisfy the assumptions of Theorem 2. The rejection sets $\mathcal{R}_k$ are defined via (9) with $p_* = 0.5$ and $\overline{\mathrm{FDP}}$ is defined via (13) with $p_* = \lambda = 0.5$. Therefore, we may apply part 2 of Theorem 2, plugging in $B = \frac{p_*}{1-\lambda} = 1$. □

3.1.3. *Interactive path.* In the *interactive setting*, p-values are split into "orthogonal" parts, with one part being used—together with side information—to determine a hypothesis ordering π and the other part being used for FDR control. AdaPT (Lei and Fithian (2018)) uses the masked p-values $g(p_j) = \min(p_j, 1 - p_j)$ and side information x_j to build up the ordering, defining a path based on (9) with $p_* = 0.5$. It then uses $\widehat{V}_{\text{preorder-sel}}$ with $p_* = \lambda = 0.5$ to construct a FDP estimate based on which the algorithm chooses a rejection set. This procedure is like Selective SeqStep, but with the ordering constructed interactively. STAR (Lei, Ramdas and Fithian (2017)), on the other hand, is the interactive analog of the accumulation test, using $p_* = 1$ and $\widehat{V}_{\text{preorder-acc}}$. It is shown that any bounded accumulation function h has a corresponding orthogonal masking function g, based on which the ordering can be constructed.

For our simultaneous FDP bounds, we use a slightly different path definition than AdaPT and STAR: we build up the path π from beginning to end, while these two methods proceed in the opposite direction. However, we do not expect this change to impact the quality of the constructed path. The path construction we consider is as follows. $\pi(1)$ is chosen based on the information $\sigma(\{x_j, g(p_j)\}_{j\in[n]})$. Once $\pi(1)$ is chosen, the corresponding p-value $p_{\pi(1)}$ is unmasked, so the information $\sigma(\{x_j, g(p_j)\}_{j\in[n]}, p_{\pi(1)})$ can be used to choose $\pi(2)$. In general, we can choose $\pi(k+1)$ in any way based on the information

$$\mathcal{G}_k \equiv \sigma\big(\{x_j, g(p_j)\}_{j\in[n]}, \{p_{\pi(j)}\}_{j\le k}\big). \tag{15}$$

Therefore, as in AdaPT and STAR, the ordering π may be built up interactively, with a human in the loop deciding the order based on $\mathcal{G}_k$. The following theorem provides FDP bounds for interactively constructed paths.

THEOREM 3. *Let π be any ordering predictable with respect to the filtration* (15), *where g is a masking function as defined below, and let*

$$\mathcal{R}_k \equiv \{\pi(j) : j \le k,\, p_{\pi(j)} \le p_*\}.$$

1. *Let h be an accumulation function bounded by B and let g is its corresponding masking function* (*see Lei, Ramdas and Fithian* (2017)). *Set $p_* = 1$, and define $\overline{\mathrm{FDP}}^B_{\text{interact-acc}} \equiv \overline{\mathrm{FDP}}^B_{\text{preorder-acc}}$. If the null p-values are independent of each other and of the nonnull p-values, and the null p-values have nondecreasing densities, then uniform bound* (5) *holds for all $a > 0$ and all $\alpha \in (0, 1)$:*

$$\Pr\{\mathrm{FDP}(\mathcal{R}_k) \le \overline{\mathrm{FDP}}^B_{\text{interact-acc}}(\mathcal{R}_k) \text{ for all } k \in [n]\} \ge 1 - \alpha.$$

2. *Fix $p_* \in (0, 1)$ and $\lambda \ge p_*$. Define $g(p) = \min(p, \frac{p_*}{1-p_*} p)$ and $\overline{\mathrm{FDP}}^B_{\text{interact-sel}} \equiv \overline{\mathrm{FDP}}^B_{\text{preorder-sel}}$. If the null p-values are independent of each other and of the nonnulls, and the null p-values are mirror-conservative* (*see Lei and Fithian* (2018)), *then uniform bound* (5) *holds for all $\alpha \in (0, 1)$:*

$$\Pr\{\mathrm{FDP}(\mathcal{R}_k) \le \overline{\mathrm{FDP}}^B_{\text{interact-sel}}(\mathcal{R}_k) \text{ for all } k \in [n]\} \ge 1 - \alpha.$$

These results are similar to the previous section's bounds, but are more subtle due to the data-dependent ordering π.

3.2. *FDP bounds for any online algorithm.* Now, we turn to FDP bounds for the online setting. In this setting, decisions about hypotheses must be made as they arrive one at a time in a stream. Moreover, the order in which hypotheses arrive might or might not be the in the experimenter's control. Therefore, nonnulls might not necessarily occur early, and further the rejection decision for the H_k must be made without knowing the outcomes of future experiments. Hence, in general, online multiple testing procedures must proceed differently from batch ones: online procedures adaptively produce a sequence of levels α_j at which to test hypotheses. Assuming for simplicity that $\pi(j) = j$, these levels define the online path:

$$\Pi_{\text{online}} \equiv (\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n, \ldots) \quad \text{where } \mathcal{R}_k \equiv \{j \le k : p_j \le \alpha_j\}. \tag{16}$$

The levels α_j are chosen based on the outcomes of past experiments, that is,

$$\alpha_{k+1} \in \mathcal{G}_k \supseteq \sigma\big(I(p_j \le \alpha_j);\, j \le k\big). \tag{17}$$

The alpha-investing procedure of Foster and Stine (2008) and follow-up works (Aharoni and Rosset (2014), Javanmard and Montanari (2018), Ramdas et al. (2017)) are built on the analogy of testing a hypothesis at level α_j as spending wealth. One pays a price to test each hypothesis, and is rewarded for each rejected hypothesis. For each of these methods, the levels α_j are adaptively constructed to ensure that the wealth always remains nonnegative. In this paper, we consider paths of the form (16) corresponding to arbitrary sequences $\{\alpha_j\}$ satisfying requirement (17), including those constructed by existing algorithms but any others as well.

Until recently, online FDR methods were formulated without reference to any $\widehat{\mathrm{FDP}}$. However, Ramdas et al. (2017) noted that LORD (Javanmard and Montanari (2018)) implicitly bounds $\widehat{\mathrm{FDP}}(\mathcal{R}_k)$ for $a_0 = 0$ and

$$\widehat{V}_{\text{online-simple}}(\mathcal{R}_k) \equiv \sum_{j=1}^{k} \alpha_j.$$

They also used this fact to design a strictly more powerful algorithm called LORD++. Moving beyond LORD++, Ramdas et al. (2018) proposed an adaptive algorithm called SAFFRON, which uses $a_0 = 0$ and

$$\widehat{V}_{\text{online-adaptive}}(\mathcal{R}_k) \equiv \sum_{j=1}^{k} \frac{\alpha_j}{1-\lambda_j} I(p_j > \lambda_j).$$

SAFFRON improves upon the LORD estimate by correcting for the proportion of nulls, making it the online analog of the Storey-BH procedure (Storey, Taylor and Siegmund (2004)). Like the levels α_j, the constants λ_j may also be chosen based on the outcomes of prior experiments.

The following theorem provides FDP bounds (4) corresponding to the above two choices for $\widehat{V}_{\text{online}}$.

THEOREM 4. *Fix $a > 0$ and let $\alpha_1, \alpha_2, \ldots$ be any sequence of thresholds predictable with respect to filtration $\mathcal{G}_k$, as in (17). Suppose the null p-values are stochastically larger than uniform conditional on the past*:

$$\Pr\{p_k \le s \mid \mathcal{G}_{k-1}\} \le s \quad \text{for each } k \in \mathcal{H}_0 \text{ and each } s \in [0,1]. \tag{18}$$

1. *Define*

$$\overline{V}_{\text{online-simple}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{a \log(1 + \frac{\log(\frac{1}{\alpha})}{a})} \cdot \left(a + \sum_{j=1}^{k} \alpha_j \right) \right\rfloor.$$

Then uniform bound (5) holds for all $\alpha \in (0,1)$:

$$\Pr\{\text{FDP}(\mathcal{R}_k) \le \overline{\text{FDP}}_{\text{online-simple}}(\mathcal{R}_k) \text{ for all } k \ge 0\} \ge 1 - \alpha.$$

2. *Let $\lambda_j \ge \alpha_j$ for all j, $\{\lambda_j\}$ be predictable with respect to $\mathcal{G}_k$, and $\sup_j \frac{\alpha_j}{1-\lambda_j} \equiv B < \infty$. Define*

$$\overline{V}^{B}_{\text{online-adaptive}}(\mathcal{R}_k) \equiv \left\lfloor \frac{\log(\frac{1}{\alpha})}{a \log(1 + \frac{1-\alpha^{B/a}}{B})} \cdot \left(a + \sum_{j=1}^{k} \frac{\alpha_j}{1-\lambda_j} I(p_j > \lambda_j) \right) \right\rfloor.$$

Then uniform bound (5) holds for all $\alpha \in (0,1)$:

$$\Pr\{\text{FDP}(\mathcal{R}_k) \le \overline{\text{FDP}}_{\text{online-adaptive}}(\mathcal{R}_k) \text{ for all } k \ge 0\} \ge 1 - \alpha.$$

For example, suppose we set $\lambda_j = 1/2$ and ran an online FDR algorithm at level $q = 0.05$. Then we would have $\alpha_j \le q$ for all j, allowing us to set $B = 2q = 0.1$. Choosing a confidence level of $\alpha = 0.1$ and additive regularization $a = 1$, we obtain $\overline{V}(\mathcal{R}_k) \equiv \lfloor 2.06 \cdot (1 + \sum_{j=1}^{k} 2\alpha_j I(p_j > 0.5)) \rfloor$.

The closest existing result to Theorem 4 is that of Javanmard and Montanari (2018) (JM). JM consider a truncated version of generalized alpha-investing rules that satisfy a uniform FDX bound like $\Pr\{\sup_k \text{FDP}_k \ge \gamma\} \le \alpha$. Their result is similar in spirit to part 1 of Theorem 4, but there are subtle differences. Their results, like most other FDX bounds, are pre hoc, meaning that given a $\gamma, \alpha \in (0,1)$, their procedure produces a sequence of rejections satisfying the FDX guarantee. Our guarantees are post hoc, meaning that they would apply to any sequence of rejections produced by any online algorithm, that may or may not have been designed for FDR or FDP control.

3.3. *A glimpse of the proof.* In this section, we present a key exponential tail inequality lemma (Lemma 1) that underlies the proofs of Theorems 2, 3 and 4. The proof of Theorem 1 requires a more involved proof technique, which we defer to the Supplementary Material (Section A), where we also show how Theorems 2, 3 and 4 follow from Lemma 1 below (Section B). We use a martingale-based proof technique that is distinct from the technique used to prove FDR control; see Section 4.2 for a comparison.

LEMMA 1. *Consider a (potentially infinite) set of hypotheses* $H_1, H_2, \dots$, *an ordering* $\pi(1), \pi(2), \dots$, *and a set of cutoffs* $\alpha_1, \alpha_2, \dots$. *Let*

$$\mathcal{R}_k \equiv \{j \le k : p_{\pi(j)} \le \alpha_j\} \quad \text{and} \quad \widehat{\mathrm{FDP}}_a(\mathcal{R}_k) \equiv \frac{a + \sum_{j \le k} h_j(p_{\pi(j)})}{|\mathcal{R}_k|},$$

where $\{h_j\}_{j \ge 1}$ *are functions on* $[0,1]$ *and the subscript* a *on* $\widehat{\mathrm{FDP}}_a$ *makes the dependence on the regularization* $a > 0$ *explicit. Suppose there exists a filtration*

$$\mathcal{F}_k \supseteq \sigma\big(\mathcal{H}_0, \{\pi(j)\}_{j \le k}, \{h_j(p_{\pi(j)}), I(p_j \le \alpha_j)\}_{j \le k, \pi(j) \in \mathcal{H}_0}\big) \tag{19}$$

such that for all $\pi(k) \in \mathcal{H}_0$, *we have*

$$\Pr\{p_{\pi(k)} \le \alpha_k \mid \mathcal{F}_{k-1}\} \le \alpha_k \quad \text{and} \quad \mathbb{E}\big[h_k(p_{\pi(k)}) \mid \mathcal{F}_{k-1}\big] \ge \alpha_k, \tag{20}$$

almost surely. Then, for each $x > 1$ *and* $a > 0$,

$$\Pr\left\{\sup_{k \ge 0} \frac{\mathrm{FDP}(\mathcal{R}_k)}{\widehat{\mathrm{FDP}}_a(\mathcal{R}_k)} \ge x\right\} \le \exp(-a\theta_x x), \tag{21}$$

where θ_x *is defined in the following four cases*:

1. *If* $h_k = h$ *for some accumulation function* h, $\alpha_k = 1$, $\pi(k)$ *is pre-specified* (*i.e., non-random*), *and* $p_{\pi(k)} \perp\!\!\!\perp \mathcal{F}_{k-1}$ *for all* $\pi(k) \in \mathcal{H}_0$, *then* θ_x *is the unique positive root of the equation*

$$\int_0^1 \exp\big(-\theta x h(u)\big)\, du = \exp(-\theta). \tag{22}$$

2. *If* $h_k = h$ *for some accumulation function* h *bounded by* B *and* $\alpha_k = 1$, *then* θ_x *is the unique positive root of the equation*

$$\exp(-\theta) + \frac{1 - \exp(-\theta x B)}{B} = 1. \tag{23}$$

3. *If* $h_k(p) = 0$ *for all* $p \le \alpha_k$, *and* $h_k(p) \le B$ *for all* k, p, *then* θ_x *is the unique positive root of the equation*

$$\exp(\theta) - \frac{1 - \exp(-\theta x B)}{B} = 1. \tag{24}$$

4. *If* $h_k(p_k) = \alpha_k$, *then* θ_x *is the unique positive root of the equation*

$$e^{\theta} = 1 + \theta x. \tag{25}$$

Let us outline the proof of the lemma. Fix any arbitrary $x > 1$ and $\theta > 0$. We first restrict our attention to only the nulls as follows:

$$\begin{aligned}&\Pr\left\{\sup_{k \ge 0} \frac{\mathrm{FDP}(\mathcal{R}_k)}{\widehat{\mathrm{FDP}}_a(\mathcal{R}_k)} \ge x\right\} \\ &\quad = \Pr\left\{\sup_{k \ge 0} \frac{V(\mathcal{R}_k)}{a + \widehat{V}(\mathcal{R}_k)} \ge x\right\}\end{aligned}$$

$$= \Pr\left\{\sum_{j=1}^{k} I(p_{\pi(j)} \le \alpha_j) I\big(\pi(j) \in \mathcal{H}_0\big) \ge ax + x \sum_{j=1}^{k} h_j(p_{\pi(j)}), \text{ for some } k \ge 0\right\}$$

$$\le \Pr\left\{\sum_{j=1}^{k} I(p_{\pi(j)} \le \alpha_j) I\big(\pi(j) \in \mathcal{H}_0\big) \ge ax + x \sum_{j=1}^{k} h_j(p_{\pi(j)}) I\big(\pi(j) \in \mathcal{H}_0\big),\right.$$

$$\left. \text{for some } k \ge 0\right\}.$$

Now, we may rearrange terms and employ the Chernoff exponentiation trick, to conclude that

$$\Pr\left\{\sup_{k\ge 0} \frac{\mathrm{FDP}(\mathcal{R}_k)}{\widehat{\mathrm{FDP}}_a(\mathcal{R}_k)} \ge x\right\}$$

$$= \Pr\left\{\sup_{k\ge 0} \exp\left(\theta\left(\sum_{j=1}^{k} [I(p_{\pi(j)} \le \alpha_j) - x h_j(p_{\pi(j)})] I\big(\pi(j) \in \mathcal{H}_0\big)\right)\right) \ge \exp(a\theta x)\right\}$$

$$\equiv \Pr\left\{\sup_{k\ge 0} Z_k \ge \exp(a\theta x)\right\}.$$

We claim that if $\theta = \theta_x$, then Z_k is a supermartingale with respect to $\mathcal{F}_k$. If this is the case, then the conclusion of the lemma would follow from the Ville (1939) maximal inequality for positive supermartingales:

$$\Pr\left\{\sup_{k\ge 0} Z_k \ge \exp(a\theta x)\right\} \le \exp(-a\theta x)\mathbb{E}[Z_0] = \exp(-a\theta x),$$

as desired. Hence, what remains is to show that in each of the four cases, the choices of θ_x make Z_k a supermartingale. To derive the FDP bounds in Theorems 2, 3 and 4. we set

$$\overline{\mathrm{FDP}}(\mathcal{R}_k) \equiv x \cdot \widehat{\mathrm{FDP}}_a(\mathcal{R}_k), \tag{26}$$

where x is chosen such that $\exp(-a\theta_x x) = \alpha$. We defer these derivations to Section B in the Supplementary Material. In fact, our definition (4) has an added floor function in the numerator, which we may add for free because the true number of false discoveries $V(\mathcal{R}_k)$ is always an integer.

4. Comparisons to work on FDR control. The paths and FDP bounds we construct are closely tied to existing FDR control algorithms. Table 2 shows each of our bounds as well as the FDR methods they are related to. In this section, we explore the relationships between our results and those already existing in the FDR literature.

4.1. *Comparing the roles of* $\widehat{\mathrm{FDP}}$. We start by recalling the definition (2) of $\widehat{\mathrm{FDP}}$. Batch FDR algorithms use this estimate of FDP to automatically choose the rejection set $\mathcal{R}^* \in \Pi$, which is done via (3). On the other hand, we use a regularized $\widehat{\mathrm{FDP}}_a$ as a building block for our confidence envelopes $\overline{\mathrm{FDP}}$ (recall definitions (4) and (26)), which the user may then inspect to choose $\mathcal{R}^*$. It is important to remark here that while our bounds are *inspired* by existing FDR algorithms, they are not intrinsically tied to the use of those procedures in any way. Indeed, we often employ the path Π of known FDR procedures, but not their stopping criterion or choice of final rejected set.

Each FDR algorithm comes with a "built-in" choice of regularization a_0. For example, the BH algorithm uses no regularization (i.e., $a_0 = 0$), while accumulation tests (Li and Barber (2017)) use $a_0 = \sup_{u\in[0,1]} h(u)$. The built-in regularizations are chosen to ensure FDR

TABLE 2
Overview of proposed FDP bounds and FDR procedures inspiring them (h denotes an accumulation function)

Ordering	p-val cutoffs	$\widehat{V}(\mathcal{R}_k)$	FDR method
sort	$p_* = 1$	$n \cdot p_{(k)}$	BH
preorder	$p_* = 1$	$\sum_{j \le k} h(p_j)$	Accumulation test
preorder	$p_* \in (0, 1)$	$\sum_{j \le k} \frac{p_*}{1-\lambda} I(p_j > \lambda)$	Selective and Adaptive SeqStep
interact	$p_* = 1$	$\sum_{j \le k} h(p_{\pi(j)})$	STAR
interact	$p_* \in (0, 1)$	$\sum_{j \le k} \frac{p_*}{1-\lambda} I(p_{\pi(j)} > \lambda)$	AdaPT
(online)	$\alpha_j \in \mathcal{G}_{j-1}$	$\sum_{j \le k} \alpha_j$	LORD, LORD++
(online)	$\alpha_j \in \mathcal{G}_{j-1}$	$\sum_{j \le k} \frac{\alpha_j}{1-\lambda_j} I(p_j > \lambda_j)$	SAFFRON, alpha-investing

control (see below). On the other hand, we consider arbitrary regularizations $a > 0$, with different regularizations leading to different constants $c(\alpha)$ and, therefore, different confidence envelopes. Different regularization parameters lead to envelopes that are tighter in different places; we have found $a = 1$ to be a good choice.

4.2. *Comparing proof techniques*. For each existing batch FDR algorithm, FDR control is established using the following martingale argument. First, the ratio $\frac{\mathrm{FDP}(\mathcal{R}_k)}{\widehat{\mathrm{FDP}}(\mathcal{R}_k)} = \frac{V(\mathcal{R}_k)}{a_0 + \widehat{V}(\mathcal{R}_k)}$ is upper bounded by a stochastic process L_k, such that L_k is a supermartingale with respect to a *backwards* filtration $\{\Omega_k\}_{k=n,\dots,1}$. Furthermore, it is shown that $\mathbb{E}[L_n] \le 1$. The choice of regularization a_0 is usually made to ensure the existence of such an L_k. Using the fact that k^* picked using rule (3) is a stopping time with respect to Ω_k, we obtain $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}(\mathcal{R}_{k^*})] \le q$ in a single line using the optional stopping theorem:

$$\frac{\mathbb{E}[\mathrm{FDP}(\mathcal{R}_{k^*})]}{q} \le \mathbb{E}\left[\frac{\mathrm{FDP}(\mathcal{R}_{k^*})}{\widehat{\mathrm{FDP}}(\mathcal{R}_{k^*})}\right] \le \mathbb{E}[L_{k^*}] \le 1.$$

This technique was first used by Storey, Taylor and Siegmund (2004) for the BH procedure, but remarkably, the other batch procedures mentioned in this paper like knockoffs, AdaPT, STAR, ordered tests and others, all implicitly use the same argument (though it was not expressed as succinctly as above), each with different L_k, Ω_k. While we also rely on a martingale argument to prove our FDP bounds (recall Section 3.3), the martingales we construct are fundamentally different: they are exponential and employ forward filtrations instead of backwards ones.

Note that the original supermartingales (L_k, Ω_k) used to prove FDR control for batch procedures can also be used to obtain tail bounds like (21), for original regularization $a = a_0$. Indeed, using Ville's maximal inequality again, we find

$$\Pr\left\{\sup_{0 \le k \le n} \frac{\mathrm{FDP}(\mathcal{R}_k)}{\widehat{\mathrm{FDP}}(\mathcal{R}_k)} \ge x\right\} \le \Pr\left\{\sup_{0 \le k \le n} L_k \ge x\right\} \le \frac{1}{x}\mathbb{E}[L_n] \le \frac{1}{x}.$$

Therefore, for each batch procedure we consider $\overline{\mathrm{FDP}}(\mathcal{R}_k) = \frac{1}{\alpha}\widehat{\mathrm{FDP}}(\mathcal{R}_k)$ is also a valid upper confidence band for FDP. Versions of this bound have been considered before in the case of BH, for example, by Robbins (1954) and Goeman et al. (2019). This implies that for all considered batch procedures, we have

$$\mathrm{Median}\left[\sup_{\mathcal{R} \in \Pi} \frac{\mathrm{FDP}(\mathcal{R})}{\widehat{\mathrm{FDP}}(\mathcal{R})}\right] \le 2.$$

However, note that the constants $c(\alpha) = \alpha^{-1}$ grow quickly as α decays. On the other hand, the constants we provide scale logarithmically, rather than linearly, in α^{-1}.

4.3. *Comparing assumptions.* This martingale argument for FDR control and the argument we employ here both require some form of independence among the p-values. Furthermore, our assumptions for each of these theorems are identical to or weaker than the ones needed to prove FDR control. For Theorem 1, we only need to make assumptions on the distribution $(p_j)_{j\in\mathcal{H}_0}$, so unlike existing proofs of FDR control for BH, we do not make any assumptions on the dependence of the nulls on the nonnulls (see Dwork, Su and Zhang (2018) for another example of such a result). In Theorem 2, we assume that the nulls are independent and stochastically larger than uniform, whereas for the original FDR control results (Barber and Candès (2015), Li and Barber (2017)) it was also required that nulls be independent of nonnulls. Furthermore, part 1 of Theorem 2 provides a FDP bound for possibly unbounded accumulation functions, whereas the original work proposing accumulation tests (Li and Barber (2017)) requires accumulation functions to be bounded. In Theorems 3 and 4, our assumptions are identical to those in the original works. Finally, we remark that the only FDR procedure which has a guarantee under dependence is BH, for which a non-martingale proof was proposed by (Benjamini and Yekutieli (2001)).

Next, we illustrate the performance of some of our bounds in simulations.

5. Numerical simulations. In this section, we compare the proposed FDP bounds to existing bounds, in the sorted and preordered settings. We also examine the effect of correlation on the proposed bounds. In all cases, we take $n = 2500$ and $\alpha = 0.1$. For the proposed bounds, we take $a = 1$.

5.1. *Sorted setting.* As discussed in the Introduction, the setting in which the most prior work has been done is when hypotheses are ordered based on p-value. In other words, we are concerned with bounds $\overline{\mathrm{FDP}}$ for the sets $\mathcal{R}_t \equiv \{j : p_j \le t\}$ such that

$$\Pr\{\mathrm{FDP}(\mathcal{R}_t) \le \overline{\mathrm{FDP}}(\mathcal{R}_t) \text{ for all } t \in [0,1]\} \ge 1-\alpha. \tag{27}$$

5.1.1. *Comparing to other explicit, finite-sample bounds.* The bounds most comparable to ours are explicit, finite-sample bounds. Two such bounds were proposed by Meinshausen and Rice (2006): $\overline{\mathrm{FDP}}(\mathcal{R}_t) \equiv \frac{\overline{V}(t)}{|\mathcal{R}_t|}$ for

$$\overline{V}_{\mathrm{Robbins}}(t) \equiv \frac{1}{\alpha}nt; \qquad \overline{V}_{\mathrm{DKW}}(t) \equiv \sqrt{\frac{n}{2}\log\frac{1}{\alpha}} + nt.$$

These bounds derive from inequalities by Robbins (1954) and Dvoretzky, Kiefer and Wolfowitz (1956), respectively. Note that inequality (27) for $\overline{V}_{\mathrm{DKW}}$ is based on the one-sided DKW inequality and is valid for $\alpha < 0.5$. Compare these to our bound, which is

$$\overline{V}_{\mathrm{sort}}(t) \equiv \frac{\log(\frac{1}{\alpha})}{\log(1+\log(\frac{1}{\alpha}))} \cdot (1+nt).$$

Note that we may add the floor function to all three of these bounds, omitted here for simplicity. By inspecting these three bounds, we see that the DKW bound is the tightest when t is large, the Robbins bound is the tightest when t is small, and our bound is the tightest in an intermediate range.

The left panel of Figure 3 compares the three bounds (floor functions included), with dotted vertical lines indicating the Bonferroni level and the nominal level, respectively. The interval between these two levels is often the most interesting for multiple testing purposes, and the proposed bound is the tightest over most of this range (in particular, it is tighter than the Robbins bound as long as $\overline{V}_{\mathrm{Robbins}}(t) \ge 2.4$). In fact, the proposed bound is not too far

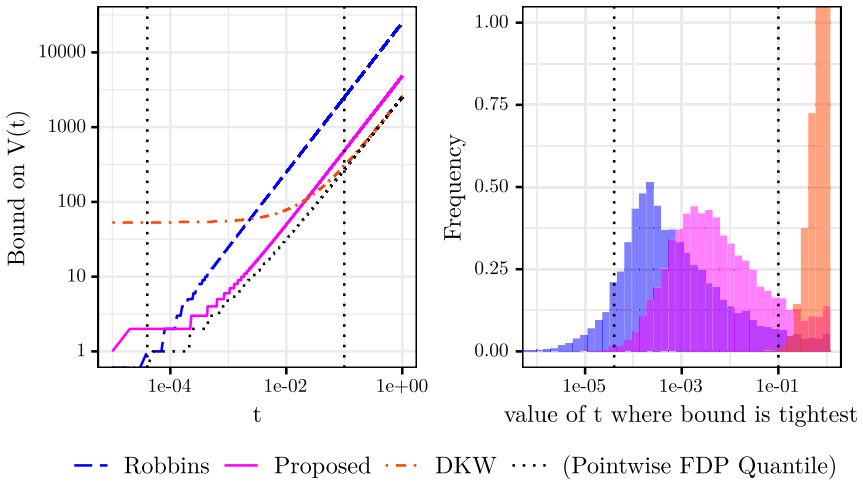

FIG. 3. *Comparing proposed bounds to two other explicit finite-sample bounds in the sorted setting. Vertical dotted lines indicate Bonferroni level α/n and nominal level α. Left panel: The three FDP bounds; the proposed bound is tighter than the other two across most of the interesting range. Right panel: histograms of t where the bound (27) is tightest.*

from the pointwise $1-\alpha$ quantile of $V(t)$, which is plotted for reference in black in the left panel. The right panel of Figure 3 shows a histogram of the value of t at which the bound (27) is tightest. We see that the majority of the time (about 87%), our bound is tightest in the interesting range.

5.1.2. *Comparison to GS bound.* As discussed in the Introduction, the GS bound is based on a suite of local tests $\{\phi_{\mathcal{R}}\}_{\mathcal{R}\in 2^{\mathcal{H}}}$. Therefore, different bounds can be obtained for different local tests. Here, we compare the proposed bound to the GS bound based on the Simes and Fisher local tests. We note that the Simes local test rejects if and only if the Robbins bound is nontrivial for any t. In fact, the GS-Simes bound is the closure of the Robbins bound and therefore dominates it (Goeman, Hemerik and Solari (2019)), so we remove the latter from consideration in this section.

Since the GS bound is not explicit, we must make the comparison by inspecting the average shape of $\overline{\mathrm{FDP}}$ on simulated data. We simulate independent test statistics $X_j \sim N(\mu_j, 1)$, where $\mu_j = \mu I(j \in \mathcal{H}_1)$ for some signal strength $\mu > 0$ and set of nonnulls $\mathcal{H}_1$. We then compute one-side p-values $p_j = 1 - \Phi(X_j)$. To cover a broad range of data-generating distributions, we consider the values $\mu = 2, 3, 4$ (weak, medium and strong signal) and $|\mathcal{H}_1| = 100, 200, 300$.

Figure 4 shows the average $\overline{\mathrm{FDP}}$ curves (over 100 repetitions) in each of the nine simulation scenarios for the proposed and GS bounds, as well as the DKW bound introduced before. For reference, the $1-\alpha$ quantile of the true FDP is also shown. We see that the GS bounds inherit the properties of their underlying local tests. The GS-Simes bound behaves like the Robbins bound: it is tightest for small rejection set sizes, yielding highly nontrivial bounds near the beginning of the path for most simulation scenarios. The GS-Fisher bound behaves the opposite way: it is tightest for large rejection set sizes, even more so than the DKW bound. Neither the GS-Fisher bound nor the DKW bound yield very informative bounds in most of the simulation settings considered. Finally, the proposed bound is an intermediate between these two extremes, yielding the tightest estimates for intermediate rejection set sizes. For the simulation settings considered, the proposed bounds are tightest in interesting regions of the path: where many rejections are made but the FDP bound is still fairly low (e.g., below 0.2).

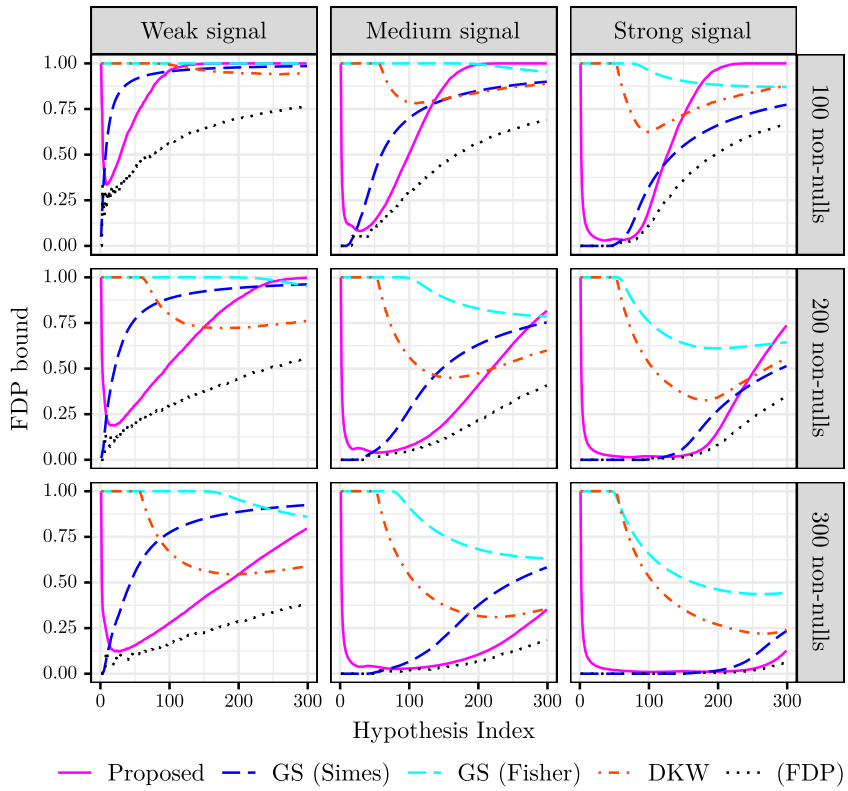

FIG. 4. *Comparing the proposed FDP bound with the GS bound (based on Simes or Fisher local tests) and the DKW bound in the sorted setting. The* $1-\alpha$ *quantile of the true FDP is also shown. The panels correspond to the three signal strengths and numbers of nonnulls. The proposed bounds are tightest in an intermediate range of rejection set sizes.*

5.1.3. *Coverage properties of FDP estimate.* The estimate $\widehat{\mathrm{FDP}}_{\mathrm{sort}}(t) = m \cdot t/|\mathcal{R}_t|$ from equation (11) and the related q-value, both proposed by Storey, Taylor and Siegmund (2004), have been shown to have asymptotic uniform coverage properties. In particular, their Theorem 6 states that for all $\delta > 0$,

$$\lim_{n\to\infty} \inf_{t\geq\delta} \{\widehat{\mathrm{FDP}}(t) - \mathrm{FDP}(t)\} \geq 0 \quad \text{with probability 1.} \tag{28}$$

At first glance, this result might suggest that there is no reason to use conservative bounds for the FDP, if asymptotically, the much smaller point estimate bounds the FDP across the entire path. However, such a conclusion is misleading. Note that the infimum in the bound (28) excludes $t \in [0, \delta)$, so for the bound to be interesting the value of δ must be small. Furthermore, the convergence becomes slower as $\delta \to 0$, so in finite samples the FDP estimate might undershoot the true FDP at some points along the path. As a counterpoint to (28), consider the following two results, holding as long as the null p-values are uniform and independent, and $\frac{|\mathcal{H}_0|}{n} \geq \epsilon > 0$:

$$\mathbb{E}\left[\sup_{t\in[0,1]} \frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)}\right] = \infty; \qquad \limsup_{n\to\infty} \sup_{t\in[\frac{c}{n},1]} \frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)} = \infty \quad \text{almost surely.}$$

The first result is due to Robbins (1954) and holds for any finite n, and the second is due to Wellner (1978) and holds for any fixed $c \geq 0$. Therefore, $\widehat{\mathrm{FDP}}$ can underestimate FDP by large factors if we remove the restriction on t or lower bound it from below by an $O(1/n)$ term.

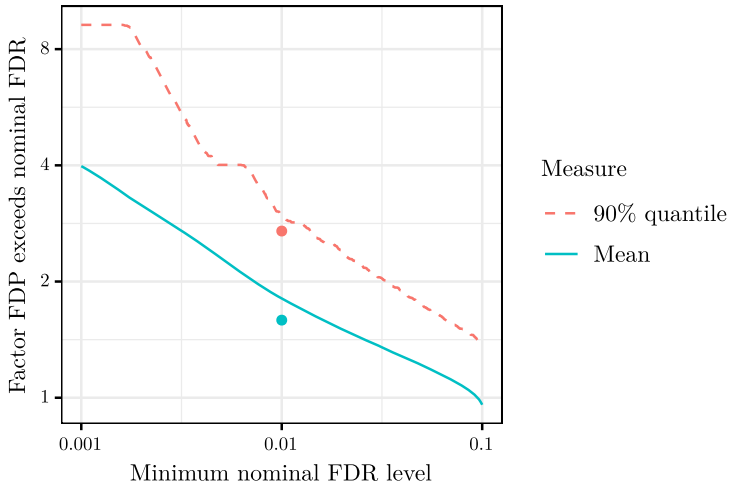

FIG. 5. *The extent to which the true FDP can exceed the nominal FDR level if BH is run for all nominal levels* $q \in [q_{\min}, 1]$. *Points correspond to running BH for the smaller prechosen set* $\mathcal{Q}_0 = \{0.01, 0.025, \ldots, 0.2\}$ *of FDR levels. We observe that some correction for exploration is necessary, whether the mean or the upper quantile of the FDP is of interest.*

We investigate via simulation whether applying BH at various target levels without correction for exploration has any consequences in finite samples. In the simulation setting from the previous section with $\mu = 4$ and $|\mathcal{H}_1| = 100$, consider applying BH using nominal levels $q \in \mathcal{Q}$ for some set $\mathcal{Q} \subseteq [0, 1]$. If $\mathrm{FDP}_{\mathrm{BH}}(q)$ denotes the realized FDP of running BH at level q, then the quantity

$$\max_{q \in \mathcal{Q}} \frac{\mathrm{FDP}_{\mathrm{BH}}(q)}{q}$$

measures the maximum extent to which the realized FDP exceeds the nominal level. In Figure 5, we show the mean and upper 90% quantile of the above quantity for $\mathcal{Q} = [q_{\min}, 1]$, with $q_{\min}$ taking a range values. The practitioner may be willing to restrict her attention to a smaller set of nominal levels chosen in advance, for example, $\mathcal{Q}_0 = \{0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2\}$. The individual points in the figure correspond to this choice.

Figure 5 shows that the true FDP can significantly exceed the nominal level, a less comforting result than (28). Note that Theorem 1 covers the case $q_{\min} = 0$ and gives a bound on the quantile of the FDP, so the corrections we introduce should be compared to the leftmost point of the red dashed curve in Figure 5. On the other hand, this figure does show that the less we allow ourselves to explore, the smaller a price needs to be paid. We see this from the monotonically decreasing trend of the curves as a function of $q_{\min}$, and from the fact that the set $\mathcal{Q}_0 \subseteq [0.01, 1]$ results in smaller factors than $[0.01, 1]$. Deriving theoretical bounds for FDP under these kinds of limitations on exploration is an interesting direction for future work. Nevertheless, even when we restrict exploration in the above ways, we see it is dangerous to take the nominal FDR level at face value as suggested by statement (28).

We note in passing that there has also been work on providing confidence envelopes such that the bound (27) holds asymptotically (for all t) as $n \to \infty$, for example, by Genovese and Wasserman (2004) and Meinshausen and Rice (2006). However, we do not review these here for the sake of brevity.

5.2. *Preordered setting.* Next, we consider the preordered setting (fixing $\pi(j) = j$ without loss of generality). Here, we fix the number of non-nulls at $|\mathcal{H}_1| = 100$ and instead vary

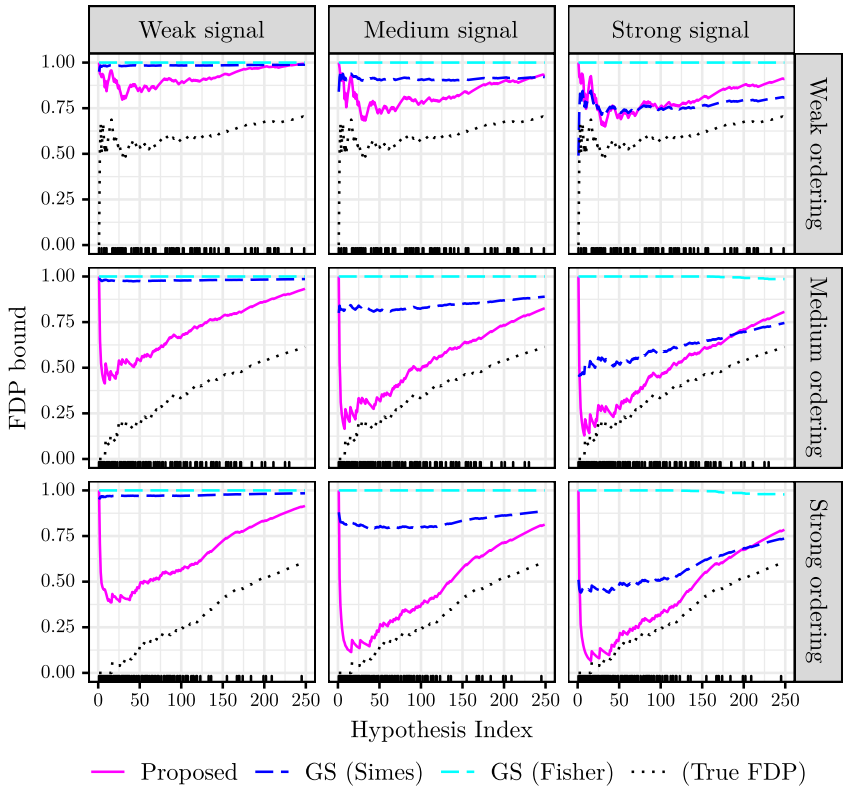

FIG. 6. *Comparing the proposed FDP bound with the GS bound based on Simes or Fisher local tests in the preordered setting. The* $1-\alpha$ *quantile of the true FDP is also shown. The panels correspond to the three signal strengths and degrees to which nonnulls occur near the beginning of the ordering. Nonnulls are shown in the rug plots at the bottom of each panel. The proposed bounds leverage the ordering information to boost power.*

the degree to which the nonnulls tend to occur near the beginning of the ordering. We sample the nonnulls without replacement from $[n]$ according to a distribution with probability mass function proportional to the density of an exponential random variable with rate θ/n. The greater θ is, the more informative the ordering is. We consider $\theta = 15, 35, 55$ (weak, medium, and strong ordering) and $\mu = 2, 3, 4$ (weak, medium and strong signal, as before). Here, the DKW and Robbins bounds are not applicable, so we only compare to GS-Simes and GS-Fisher. We apply our bound based on $\widehat{V}_{\text{preorder-acc}}$, with accumulation function $h(p) = \frac{1}{1-\lambda} I(p > \lambda)$ with $\lambda = 0.1$. We use the definition (2.) of $\overline{\text{FDP}}^{B}_{\text{preorder-acc}}$.

Figure 6 shows the results. We see that the proposed bound effectively leverages the ordering information to obtain tighter FDP bounds than the GS-based methods. Predictably, the stronger the ordering information, the greater the advantage of our bound. Consistent with the previous simulation, GS-Simes outperforms GS-Fisher; the latter bound is nearly trivial for all simulation settings. Of course, an interesting direction of future work is to derive tighter GS-style bounds for settings with prior information.

5.3. *The effect of correlation.* Finally, note that all our FDP bounds rely on some notion of independence among the p-values. Many of the FDR procedures considered here also only have guarantees under independence, though BH is a notable exception. Aside from online testing applications, independent p-values are hard to come by in practice, so more robust guarantees are necessary. BH is known to control FDR under the PRDS criterion (Benjamini and Yekutieli (2001)), a form of positive dependence that contains no information about the *strength* of the dependence. However, it is known that while the mean of FDP might not

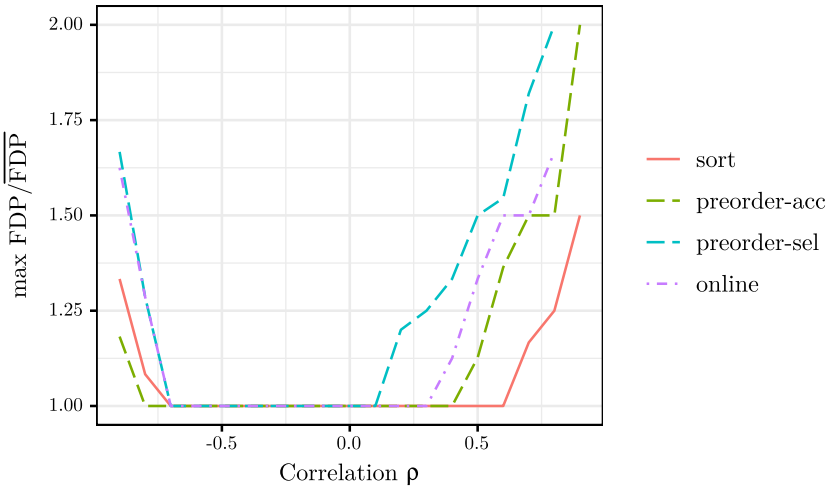

FIG. 7. *The extent to which FDP can exceed* $\overline{\text{FDP}}$ *for the proposed bounds under* p*-value correlation, generated from an AR*(1) *model parameterized by* ρ*. The bounds are more tolerant of negative than positive correlation.*

change much as dependence increases, the variance of the FDP will increase (Owen (2005), Efron (2010)). Hence, high-probability bounds on FDP under dependence are likely to use criteria other than PRDS to capture this dependence.

In this section, we use simulations to examine the extent to which our bounds continue to hold in the presence of p-value correlation. To model correlation, we draw the test statistics X_j from an AR(1) process parameterized by correlation $\rho = -0.9, -0.8, \ldots, 0.8, 0.9$. We consider four representative settings: the sorted setting from Theorem 1 and Section 5.1, the preordered setting with $p_* = 1$ from Theorem 2, part 1 and Section 5.2, the preordered setting with $p_* = \lambda = 0.1$ from Theorem 2, part 2, and the online setting with $\alpha_j = 0.05$ for all j from Theorem 4, part 1. For each setting and each value of ρ, we compute the $1 - \alpha$ quantile of $\max_k \text{FDP}(\mathcal{R}_k)/\overline{\text{FDP}}(\mathcal{R}_k)$, the maximum extent to which FDP can exceed our bound. We operate under the global null, since this is the worst case scenario.

Figure 7 shows the simulation results. Reassuringly, all curves pass through 1 at $\rho = 0$, the independent case covered by our theorems. We see that different bounds have different tolerances for correlation, but negative correlation is tolerated better than positive correlation. All bounds continue to hold for $\rho \in [-0.7, 0.1]$. The bound in the sorted setting is particularly robust, continuing to be valid for $\rho \in [-0.7, 0.6]$. Nevertheless, it is not surprising that all the bounds are no longer valid once the correlation becomes strong enough. Indeed, under strong correlations the variability of the FDP necessitates more conservative bounds. We leave the extension of our results to the correlated setting for future work.

6. Conclusion. In this paper, we establish a novel bridge between the realms of FDR control and simultaneous FDP control. While FDR procedures rely on estimates of the FDP to choose one rejection set from a path, we repurpose these estimates to obtain closed form simultaneous bounds on the FDP that are valid across the entire path with high probability. These novel bounds allow for the kind of simultaneous inference proposed by Goeman and Solari (2011), where users can obtain FDP bounds on rejection sets they choose after exploring the data. They offer added versatility, applying in the structured, regression and online settings; in Section 5, we found that our bounds effectively leverage side information to boost power.

Like any other simultaneous inference methodology, the bounds we provide must necessarily be conservative at certain points along the path. This reflects the fundamental trade-off

between exploration and inference: allowing more flexibility to explore necessitates conservative corrections for inference to remain valid. In applications where analytical flexibility is important, however, this price may be worth paying. By augmenting the recent knockoffs analysis of the UK Biobank data set (Sesia et al. (2019)) with simultaneous FDP bounds, we saw in Section 2 how much extra freedom we gained to find a biologically meaningful set of associations between genomic regions and human traits. While there might be a price in power as compared to the usual FDR analysis for other datasets, for this real-world dataset we still obtained meaningful FDP bounds for large rejection sets. Having said that, we do not necessarily advocate for employing simultaneous inference in all situations; indeed, FDR or FDX control at prespecified levels may well be the right analysis choice in a variety of applications.

Figure 5 illustrates the aforementioned trade-off between exploration and inference, and suggests that restricting the collection of rejection sets allowed to be explored can reduce the price paid for exploration. Studying this trade-off may lead to interesting future work. In addition to the fact that scientists may in practice explore fewer rejection sets than the guarantee covers, the rejection set they ultimately choose is likely not going to be the worst one in terms of FDP. Therefore, the worst-case bounds considered in this paper have this inherent degree of looseness. However, this looseness seems difficult or impossible to address theoretically.

Recently, Goeman, Hemerik and Solari (2019) explored the question of optimality among simultaneous inference procedures, proposing a natural admissibility criterion. In addition, these authors proved that only closed testing procedures, that is, those of the kind proposed by Goeman and Solari (2011), are admissible. Given any simultaneous inference procedure, like those proposed here, they showed how to improve the procedure by "closing" it. From this perspective, our results can be viewed as building blocks from which to construct more sophisticated closed testing based procedures. It is not always the case that a closed testing procedure can be implemented in polynomial time, however, so it is still not clear which simultaneous bounds are dominated by other *computationally efficient* bounds.

Our results may be employed in other contexts as well. When multiple groupings of hypotheses are of interest, as considered previously by Barber and Ramdas (2017), Katsevich and Sabatti (2019), Ramdas et al. (2019), our results can give simultaneous FDP bounds with respect to each grouping. As pointed out to us by a referee, our bounds may also be used to construct new tests of the global null. Moreover, following Meinshausen and Rice (2006), our uniform bounds can also be used to estimate the null proportion among a set of hypotheses. Exploring these consequences of the proposed bounds is an interesting direction for future work.

Finally, the proof technique we developed in this paper is versatile enough to cover a large portion of the currently available FDR procedures. Importantly, this includes the knockoffs procedure for variable selection in high dimensions. Like Genovese and Wasserman (2004), we employ a stochastic process approach to analyze the FDP. However, while GW's bounds are asymptotic, we have used martingale arguments instead to obtain tight, nonasymptotic bounds. Perhaps these proof techniques may be extended further to apply to other multiple testing scenarios as well.

Acknowledgments. E.K. thanks Chiara Sabatti for her generous and valuable feedback on this work and on the manuscript itself, as well as David Siegmund, Emmanuel Candes, Anya Katsevich and Michael Sklar for helpful discussions. A.R. acknowledges fruitful conversations with Ohad Feldheim, Jim Pitman and Jon Wellner about the BH empirical process. Both authors are also grateful for the insightful comments of the referees and Associate Editor. We acknowledge the UK Biobank Resource (application 27837) for the data used in

Section 2, and Matteo Sesia for making public the summary statistics from the knockoffs analysis of this data (Sesia et al. (2019)).

E.K.'s work was supported by the Fannie and John Hertz Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings" (DOI: 10.1214/19-AOS1938SUPP; .pdf). Proofs of all theorems.

REFERENCES

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 https://doi.org/10.1111/rssb.12048

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25.

BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2016). Uniformly valid confidence intervals post-model-selection. Available at arXiv:1611.01043.

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

BARBER, R. F. and RAMDAS, A. (2017). The *p*-filter: Multilayer false discovery rate control for grouped hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1247–1268. MR3689317 https://doi.org/10.1111/rssb.12218

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

BENJAMINI, Y. and LIU, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82** 163–170. MR1736441 https://doi.org/10.1016/S0378-3758(99)00040-3

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.* **48** 1281–1303. MR4124323 https://doi.org/10.1214/19-AOS1847

BYCROFT, C., FREEMAN, C., PETKOVA, D., BAND, G., ELLIOTT, L. T., SHARP, K., MOTYER, A., VUKCEVIC, D., DELANEAU, O. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562** 203.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864 https://doi.org/10.1214/aoms/1177728174

DWORK, C., SU, W. J. and ZHANG, L. (2018). Differentially private false discovery rate control. Available at arXiv:1807.04209v1.

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

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 https://doi.org/10.1111/j.1467-9868.2007.00643.x

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 https://doi.org/10.1111/rssb.12122

GAVRILOV, Y., BENJAMINI, Y. and SARKAR, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.* **37** 619–629. MR2502645 https://doi.org/10.1214/07-AOS586

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283

GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339

GOEMAN, J., HEMERIK, J. and SOLARI, A. (2019). Only closed testing procedures are admissible for controlling false discovery proportions. Available at arXiv:1901.04885.

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

GOEMAN, J. J., MEIJER, R. J., KREBS, T. J. P. and SOLARI, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106** 841–856. MR4046036 https://doi.org/10.1093/biomet/asz041

HEMERIK, J., SOLARI, A. and GOEMAN, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649. MR3992394 https://doi.org/10.1093/biomet/asz021

JAVANMARD, A. and MONTANARI, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* **46** 526–554. MR3782376 https://doi.org/10.1214/17-AOS1559

KATSEVICH, E. and RAMDAS, A. (2020). Supplement to "Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings." https://doi.org/10.1214/19-AOS1938SUPP.

KATSEVICH, E. and SABATTI, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* **13** 1–33. MR3937419 https://doi.org/10.1214/18-AOAS1185

KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., CAI, J., GEORGE, E. I. and ZHAO, L. (2020). Valid post-selection inference in model-free linear regression. *Ann. Statist.* To appear.

LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631 https://doi.org/10.1214/009053605000000084

LEI, L. and FITHIAN, W. (2016). Power of ordered hypothesis testing. In *International Conference on Machine Learning* 2924–2932.

LEI, L. and FITHIAN, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 649–679. MR3849338 https://doi.org/10.1111/rssb.12253

LEI, L., RAMDAS, A. and FITHIAN, W. (2017). STAR: A general interactive framework for FDR control under structural constraints. Available at arXiv:1710.02776.

LI, A. and BARBER, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Statist. Assoc.* **112** 837–849. MR3671774 https://doi.org/10.1080/01621459.2016.1180989

MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M. and BEJERANO, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28** 495–501. https://doi.org/10.1038/nbt.1630

MEINSHAUSEN, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Stat.* **33** 227–237. MR2279639 https://doi.org/10.1111/j.1467-9469.2005.00488.x

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 https://doi.org/10.1214/009053605000000741

OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 411–426. MR2155346 https://doi.org/10.1111/j.1467-9868.2005.00509.x

RAMDAS, A., YANG, F., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems*.

RAMDAS, A., ZRNIC, T., WAINWRIGHT, M. and JORDAN, M. (2018). SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the* 35*th International Conference on Machine Learning* 4286–4294.

RAMDAS, A. K., BARBER, R. F., WAINWRIGHT, M. J. and JORDAN, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.* **47** 2790–2821. MR3988773 https://doi.org/10.1214/18-AOS1765

ROBBINS, H. (1954). A one-sided confidence interval for an unknown distribution function. *Ann. Math. Stat.* **25** 409.

SESIA, M., SABATTI, C. and CANDÈS, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. MR3912377 https://doi.org/10.1093/biomet/asy033

SESIA, M., KATSEVICH, E., BATES, S., CANDÈS, E. and SABATTI, C. (2019). Multi-resolution localization of causal variants across the genome. *BioRxiv*. https://doi.org/10.1101/631390

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766 https://doi.org/10.1111/j.1467-9868.2004.00439.x

VAN DER LAAN, M. J., DUDOIT, S. and POLLARD, K. S. (2004). Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives.

VILLE, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars, Paris. MR3533075

WELLNER, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. Verw. Gebiete* **45** 73–88. MR0651392 https://doi.org/10.1007/BF00635964