Submitted to the Annals of Statistics
arXiv: arXiv:1703.06222

A UNIFIED TREATMENT OF MULTIPLE TESTING WITH PRIOR KNOWLEDGE USING THE P-FILTER

By Aaditya Ramdas*, Rina F. Barber†,
Martin J. Wainwright*, and Michael I. Jordan*

*University of California, Berkeley** *and University of Chicago*†

A significant literature studies ways of employing prior knowledge to improve power and precision of multiple testing procedures. Some common forms of prior knowledge may include (a) *a priori* beliefs about which hypotheses are null, modeled by non-uniform prior weights; (b) differing importances of hypotheses, modeled by differing penalties for false discoveries; (c) multiple arbitrary partitions of the hypotheses into known (possibly overlapping) groups, indicating (dis)similarity of hypotheses; and (d) knowledge of independence, positive or arbitrary dependence between hypotheses or groups, allowing for more aggressive or conservative procedures. We present a unified algorithmic framework called `p-filter` for global null testing and false discovery rate (FDR) control that allows the scientist to incorporate all four types of prior knowledge (a)–(d) simultaneously, recovering a wide variety of common algorithms as special cases.

1. Introduction. Multiple hypothesis testing is both a classical and highly active research area: it dates back (at least) to an initially unpublished 1953 manuscript by Tukey entitled "the problem of multiple comparisons" [40]. Given a large set of null hypotheses, multiple testing commonly deals with deciding which subset to reject, while guaranteeing some notion of control on the number of false rejections. It is of practical importance to incorporate different forms of prior knowledge into existing multiple testing procedures; such prior knowledge can yield improvements in power and precision, and can also provide more interpretable results. The focus of this paper is on methods that control the False Discovery Rate (FDR) or test the global null (GN) hypothesis while incorporating any number of the following considerations: (a) using prior weights, (b) using penalty weights, (c) partitioning the hypotheses into groups, (d) incorporating knowledge of the dependence structure within the data, including options such as estimating and adapting to the unknown number of nulls under independence, or reshaping rejection thresholds to preserve error control guarantees in the

MSC 2010 subject classifications: Primary 62J15, 60G10; secondary 62F03
Keywords and phrases: multiple testing, false discovery rate, prior knowledge, Simes, Benjamini-Hochberg-Yekutieli, adaptivity, group FDR

presence of arbitrary dependence. It is a challenge to incorporate all of these forms of structure while maintaining internal consistency (coherence and consonance) among the pattern of rejections and acceptances, and most existing work has managed to simultaneously employ only one or two of the four considerations (a), (b), (c), (d). We present a general unified framework for integrating all four, while performing a GN test or controlling FDR. This framework allows scientists to mix and match techniques, and use multiple different forms of prior knowledge simultaneously. Our framework simplifies the analysis of existing procedures, and generalize the conditions under which they are known to work. We now outline our contributions.

Outline. Rather than presenting the most general version of our framework at the outset, we introduce the framework gradually, beginning with simple weighting schemes and then introducing more tools and notation as needed to encompass more complex schemes. Since all types of prior information considered in this paper have already been motivated in applied settings, our focus is on the conceptual and mathematical aspects.

- **Assumptions.** In Section 2, we formalize the range of settings studied, in terms of the assumptions on the underlying p-values that provide the data for our multiple testing procedures—marginally, the p-values are assumed to be super-uniform, and jointly, we consider the three settings of (a) independence, (b) positive dependence, or (c) arbitrary dependence.
- **Generalized BHY procedures.** We discuss the use of reshaping functions, as introduced by Blanchard and Roquain [9], to guard against arbitrary dependence. We use these to describe a set of procedures that generalize the classical Benjamini-Hochberg (BH) procedure [3] and the Benjamini-Yekutieli (BY) procedure [7]; we thus refer to them as "generalized BHY procedures." We develop a general *super-uniformity lemma* that is at the core of FDR control proofs. As a first application, we show that it yields a short proof of FDR control for generalized BHY procedures under all three settings of dependence.
- **Prior- and/or penalty-weighted BHY.** In Section 4, we discuss two natural ways to incorporate weights—prior-weighted p-values and penalty-weighted FDR—and we describe a procedure that uses both prior and penalty weights simultaneously. Using the super-uniformity lemma, we show that this procedure controls (weighted) FDR under all three settings of dependence. This immediately implies old and new results for algorithms studied by Genovese et al. [16], Benjamini and Hochberg [4], and Blanchard and Roquain [9].
- **Adaptive, prior and/or penalty-weighted BH.** In Section 5, we present so-called "adaptive" procedures for independent p-values, which

improve power by incorporating an estimate of the proportion of nulls, as first proposed by Storey et al. [37, 38]. We then propose a novel adaptive procedure, that can employ both prior and penalty weights, as well as null proportion estimates. Using a new *inverse-binomial lemma*, we succinctly prove that our adaptive, doubly-weighted BH procedure controls (weighted) FDR under independence.

- **Grouped, adaptive, doubly-weighted procedures.** In Section 6, we tackle the problem of controlling the *group-level FDR* for a single user-defined partition of the hypotheses, when the group-level p-values may or may not be formed from the elementwise p-values. We particularly discuss one way to achieve this based on the Simes p-value, first showing that the prior-weighted Simes test studied by Hochberg and Liberman [19] is valid even under positive dependence, and can be reshaped to work under arbitrary dependence. Then, using a novel *group super-uniformity lemma*, we demonstrate how to use our novel weighted Simes tests in conjunction with doubly-weighted adaptive BH procedures to control the (weighted) group FDR.
- **A unifying multilayer, grouped, adaptive, doubly-weighted procedure.** In Section 7, we generalize the previous section to allow the use of multiple arbitrary user-defined partitions of hypotheses that are possibly non-hierarchically arranged relative to each other. In the process, we develop an algorithmic framework called `p-filter`, extending earlier work of Barber and Ramdas [1]. `p-filter` allows the use of any or all of the following features and extensions: different prior and penalty weights for every group in every partition; null-proportion adaptivity (for all partitions whose groups are known to be independent); incomplete partitions that contain a "leftover set"; overlapping groups within a partition; and reshaping functions to guard against arbitrary dependence. This algorithmic framework guarantees *internal consistency*, meaning that the rejected hypotheses and groups are always in mutual agreement. We combine the super-uniformity lemma with novel proof techniques to prove that `p-filter` guarantees (weighted) group FDR control for all partitions simulataneously, under all three forms of dependence.

Sometimes, we reprove existing results in the literature, usually for the purposes of illustrating a novel proof technique; in such cases, we label our results as propositions. In other cases, when results are new and not found elsewhere in the literature, we label them as theorems. Table 1 summarizes our contributions relative to some related work. Given space limitations, we only list a few related references, and defer more detailed discussions of related work to the point in the paper when those settings are discussed.

Form of incorporated structure	prior work	this paper
none (False Discovery Rate)	■ Benjamini and Hochberg [3] ● ● ● Benjamini and Yekutieli [7]	● ● ● Proposition 1
prior weights (FDR)	■ Genovese et al. [16]	● ● ● Proposition 2
penalty weights (FDR)	■ Benjamini and Hochberg [4] ■ ● ● Finos and Salmaso [13]	● ● ● Proposition 2
prior and/or penalty weights (FDR)	■ ● ● Blanchard and Roquain [9]	● ● ● Proposition 2
null proportion adaptivity (FDR)	● Storey et al. [38]	● Proposition 3
prior and/or penalty weighted null proportion adaptivity (FDR)	—	● Theorem 5.1
none (Global Null)	● Simes [33] ● ● ● Benjamini and Yekutieli [7]	● ● ● Proposition 4
prior weights (Global Null)	● Hochberg and Liberman [19]	● ● ● Proposition 4
single partition into groups (FDR)	● Barber and Ramdas [1]	● ● ● Proposition 5
multiple arbitrary partitions (FDR)	■ ● Barber and Ramdas [1]	● ● ● Theorem 7.1
multiple possibly incomplete arbitrary partitions of possibly overlapping groups + prior and/or penalty weights + group null proportion adaptivity (FDR)	—	● ● ● Theorem 7.1

TABLE 1

This table summarizes our contributions relative to other known results regarding step-up procedures for multiple testing The traffic lights ● , ● and ● represent results that were proved under independence (easy), positive dependence (intermediate), and arbitrary dependence (hardest). In the top box, ■ indicates that under uniformity, the FDR *bound was proved with inequality instead of strict equality. In the bottom box, ■ indicates that null-proportion adaptivity was not provided, and also the group-level p-values were constrained to be Simes p-values.*

2. Background. We begin by presenting background on multiple testing and FDR control, overviewing common assumptions in the literature, and providing a discussion of the Benjamini-Hochberg procedure. Before doing so, it is helpful to introduce some basic definitions and notation. For a pair of vectors $x, y \in [0,1]^n$, we use the notation $x \preceq y$ to mean that $x \leq y$ in the orthant ordering, i.e., $x_i \leq y_i$ for all $i \in \{1, \dots, n\}$.

DEFINITION 1 (Nondecreasing sets and functions). A set $\mathcal{D} \subseteq [0,1]^n$ is said to be *nondecreasing* if $x \in \mathcal{D}$ implies $y \in \mathcal{D}$ for all $y \succeq x$. We say that a function $f : [0,1]^n \mapsto [0, \infty)$ is *nonincreasing*, if $x \preceq y$ implies $f(x) \geq f(y)$.

Finally, in order to handle the ratio "$\frac{0}{0}$" that often arises in FDR control results, we adopt the "dotfraction" notation

$$\frac{a}{\dddot{b}} := \begin{cases} \frac{a}{b}, & \text{if } a \neq 0, b \neq 0, \\ 0, & \text{if } a = b = 0, \\ 0, & \text{if } a = 0, b \neq 0, \\ \text{undefined} & \text{if } a \neq 0, b = 0. \end{cases} \tag{2.1}$$

Dotfractions behave like fractions whenever the denominator is nonzero. We formally derive properties of dotfractions in detail in Appendix G.

2.1. *False Discovery Rate (FDR).* In the standard multiple testing setup, we are given a set of n different p-values, denoted by the random vector $P \in [0,1]^n$. Each p-value corresponds to a different null hypothesis, and we let $\mathcal{H}^0 \subseteq [n]$ denote the subset of true null hypotheses. Our goal is to reject some subset of the null hypotheses—or in other words, to select some subset of $\{1, \dots, n\}$ as our discoveries—while at the same retaining control over the number of false discoveries. Consider any algorithm that, based on the observed vector P of p-values, chooses a subset of hypotheses to reject; this random subset is denoted by $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(P)$. In their seminal work, Benjamini and Hochberg [3] proposed to measure performance of such algorithms by a quantity called the *False Discovery Rate* (FDR), and gave a simple procedure to provably control it.[1]

The FDR is defined in terms of the *False Discovery Proportion*, given by $\mathrm{FDP} := \frac{|\mathcal{H}^0 \cap \widehat{\mathcal{S}}|}{\dddot{|\widehat{\mathcal{S}}|}}$, where $|\mathcal{H}^0 \cap \widehat{\mathcal{S}}|$ is number of *false discoveries*—that null hypotheses that are true, and are incorrectly rejected—whereas $|\widehat{\mathcal{S}}|$ is the

[1] As a historical footnote, the same procedure was proposed by Eklund [11], albeit as a heuristic without any formal guarantees [12, 32].

total number of discoveries, meaning null hypotheses that are rejected. The FDR is the expectation of this random ratio:

$$\text{FDR} := \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \widehat{\mathcal{S}}|}{|\widehat{\mathcal{S}}|}\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} \mathbf{1}\left\{i \in \widehat{\mathcal{S}}\right\}}{\sum_i \mathbf{1}\left\{i \in \widehat{\mathcal{S}}\right\}}\right]. \tag{2.2}$$

Our goal is to construct procedures that control FDR at some target level $\alpha \in (0,1)$, that is to guarantee that FDR $\leq \alpha$.

2.2. *P-value dependence assumptions.* We assume that the marginal distribution of each null p-value is *super-uniform*, meaning that it is stochastically dominated by the uniform distribution. More precisely, for any index $i \in \mathcal{H}^0$, we assume that

$$\Pr\{P_i \leq t\} \leq t \quad \text{for all } t \in [0,1]. \tag{2.3}$$

Of course, uniformly distributed p-values trivially satisfy this condition. We use the phrase *under uniformity* to describe the situation in which the null p-values are marginally exactly uniform. If this phrase is not employed, it is understood that the null p-values are marginally super-uniform.

Regarding assumptions on the joint distribution of p-values, three possible kinds of dependence will be considered in this paper: independence, positive dependence or arbitrary dependence. In the independent setting, all p-values are assumed to be mutually independent. The second possibility is that of "positive dependence" as formalized by *Positive Regression Dependence on a Subset* condition, or PRDS for short:

DEFINITION 2 (PRDS). We say that the vector P satisfies PRDS if for any null index $i \in \mathcal{H}^0$ and nondecreasing set $\mathcal{D} \subseteq [0,1]^n$, the function $t \mapsto \Pr\{P \in \mathcal{D} \mid P_i \leq t\}$ is nondecreasing over $t \in (0,1]$.

The original positive regression dependence assumption as introduced by Lehmann [24] as well as the PRDS assumption first made by Benjamini and Yekutieli [7] both had $P_i = t$ instead of $P_i \leq t$ in the definition, but one can prove that both conditions are essentially equivalent.

The PRDS condition holds trivially if the p-values are independent, but also allows for some amount of "positive" dependence. Let us consider a simple example to provide some intuition. Let $Z = (Z_1, \ldots, Z_n)$ be a multivariate Gaussian vector with covariance matrix Σ; the null components correspond to Gaussian variables with zero mean. Letting Φ be the CDF of a standard Gaussian, the vector $P = (\Phi(Z_1), \ldots, \Phi(Z_n))$ is PRDS on

P_i for every index i if and only if all entries of the covariance matrix Σ are non-negative. See Benjamini and Yekutieli [7] for additional examples of this type. It should be noted the PRDS assumption is closely related to the assumption of multivariate total positivity of order 2 (MTP2), as studied by Karlin and Rinott [22]. Since MTP2 implies PRDS, all the results in this paper that hold under PRDS also immediately also hold under MTP2.

3. Generalized Benjamini-Hochberg-Yekutieli procedures. We begin by discussing the classical Benjamini-Hochberg procedure. Given a vector $P \in [0,1]^n$, we let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$ denote the associated vector of order statistics. Given these order statistics, the Benjamini-Hochberg procedure (BH) [3] with target FDR level α rejects the smallest $\widehat{k}$ p-values, where

$$\widehat{k} = \widehat{k}_\alpha(P) := \max\left\{k : P_{(k)} \leq \frac{\alpha \cdot k}{n}\right\}, \tag{3.1}$$

with that convention that $\widehat{k} = 0$ if the set is empty. That is, after choosing this value $\widehat{k}$, the set of rejections is $\widehat{\mathcal{S}} = \left\{i : P_i \leq \frac{\alpha \cdot \widehat{k}}{n}\right\}$, which by definition of $\widehat{k}$, must contain exactly $\widehat{k}$ many hypotheses. Benjamini and Hochberg [3] prove that this procedure controls the FDR at the desired level, that is, $\mathrm{FDR} \leq \alpha$ for this rejection set $\widehat{\mathcal{S}}$, under the assumption that the p-values are mutually independent. This result was shown to hold also in the positive dependence setting by Benjamini and Yekutieli [7].

When no assumptions are made about the joint distribution of p-values, however, it is conceivable that FDR-controlling procedures must be guarded while proclaiming a discovery. Benjamini and Yekutieli [7] originally proposed running the BH procedure at the modified threshold $\frac{\alpha}{\sum_{i=1}^n \frac{1}{i}}$ in order to maintain FDR control under arbitrary dependence. We now introduce a family of generalizations of this modification, which we will refer to as "generalized BY procedures" or simply "BY procedures" throughout the paper. In order to do so, we first need the notion of a *reshaping* function β, as introduced by Blanchard and Roquain [9].

Definition 3 (Reshaping). For any fixed probability measure ν on $[0,\infty)$, we define the reshaping function $\beta = \beta_\nu$ as

$$\beta_\nu(k) = \int_0^k x \; \mathsf{d}\nu(x).$$

Fixing a reshaping function β, then, the generalized Benjamini-Yekutieli (BY) procedure with target FDR level α makes

$$\widehat{k} = \widehat{k}_\alpha(P) := \max\left\{k : P_{(k)} \leq \frac{\alpha \cdot \beta(k)}{n}\right\} \tag{3.2}$$

rejections, again with that convention that $\widehat{k} = 0$ if the set is empty. The rejection set is now given by $\widehat{\mathcal{S}} = \left\{i : P_i \leq \frac{\alpha \cdot \beta(\widehat{k})}{n}\right\}$, which again contains exactly $\widehat{k}$ many hypotheses. Noting that $\beta(k) = \beta_\nu(k) \leq k$ for any probability measure ν, thus requiring that the k lowest p-values fall below a lower (more strict) threshold, we can immediately see that for any choice of β, this generalized BY procedure is more conservative than the BH procedure (if run at the same target FDR level α).

As an example, we can see that Benjamini and Yekutieli [7]'s original proposal, i.e. running the BH procedure at the modified threshold $\frac{\alpha}{\sum_{i=1}^n \frac{1}{i}}$, is equivalent to running the BY procedure with the reshaping function

$$\beta_{BY}(k) = \frac{k}{\sum_{i=1}^n \frac{1}{i}}. \tag{3.3}$$

We now summarize some known properties of the BH and BY procedures.

PROPOSITION 1. *The BH procedure* (3.1) *and generalized BY procedures* (3.2) *have the following properties:*

(a) Under independence and uniformity, the BH *method has* $\text{FDR} = \alpha\frac{|\mathcal{H}^0|}{n}$.
(b) Under positive dependence, the BH *method has* $\text{FDR} \leq \alpha\frac{|\mathcal{H}^0|}{n}$.
(c) Under arbitrary dependence, the generalized BY *procedure with any reshaping function* β *yields* $\text{FDR} \leq \alpha\frac{|\mathcal{H}^0|}{n}$.

Statement (a) was proven with inequality by Benjamini and Hochberg [3]. Benjamini and Yekutieli [7] proved statement (a) with equality, as well as proving statement (b), and statement (c) for the specific reshaping function (3.3). For an arbitrary reshaping function β, statement (c) was proven by Blanchard and Roquain [9].

3.1. *The role of reshaping.* Naturally, there is no universally optimal choice of ν, and different choices of base probability measure ν (and hence β) lead to different behaviors of the associated procedures. A typical choice of the measure ν is a discrete distribution with support on $\{1, \ldots, n\}$; such a choice is sensible since the number of rejections is always an integer. Indeed,

as noted by Blanchard and Roquain [9], any continuous measure ν can be replaced by the discrete measure

$$\nu'(r) = \begin{cases} \nu((r-1, r]) & \text{if } r < n \\ \nu((n-1, \infty)) & \text{if } r = n \end{cases}$$

to yield a reshaping function $\beta_{\nu'}$ that is strictly larger than β_ν on the relevant integer range, leading to a more powerful procedure.

In contrast with the discrete distributions considered in past work, this paper also considers the case of continuous measures ν. Continuous measures are of interest in the context of the penalty-weighted procedures to be considered in the sequel, where the function β corresponds to the total weight assigned to subsets of the hypotheses. As opposed to the count, this total weight can be fractional, so that one can no longer reduce to the case of discrete measures ν, as we did above.

Many different examples of reshaping functions, and their connections to other formulations of multiple testing methods, can be found in the literature, see e.g. Blanchard and Roquain [9], Sarkar [30, 31].

3.2. *Alternative derivation of BH and BY procedures.* There is an alternate way to motivate the BH and BY procedures, which will be useful for designing algorithms later in this paper. Let $\mathrm{FDP}(t)$ denote the FDP of a procedure that rejects all p-values less than or equal to t, and note that

$$\begin{aligned} \mathrm{FDP}(t) = \frac{|\mathcal{H}^0 \cap \widehat{\mathcal{S}}|}{|\widehat{\mathcal{S}}|} &= \frac{\sum_{i \in \mathcal{H}^0} \mathbf{1}\{P_i \le t\}}{\sum_i \mathbf{1}\{P_i \le t\}} \\ &\approx \frac{|\mathcal{H}^0| \cdot t}{\sum_i \mathbf{1}\{P_i \le t\}} \le \frac{n \cdot t}{\sum_i \mathbf{1}\{P_i \le t\}} =: \widehat{\mathrm{FDP}}(t), \end{aligned} \tag{3.4}$$

where the approximation $\sum_{i \in \mathcal{H}^0} \mathbf{1}\{P_i \le t\} \approx |\mathcal{H}^0| \cdot t$ follows from the observation that $\mathbb{E}\left[\sum_{i \in \mathcal{H}^0} \mathbf{1}\{P_i \le t\}\right] \le |\mathcal{H}^0| \cdot t$—that is, we expect roughly $|\mathcal{H}^0| \cdot t$ many of the null p-values to be $\le t$. In order to guard against unexpected dependence structures, we would reshape the denominator of $\widehat{\mathrm{FDP}}(t)$ so as to undercount rejections, replacing it by $\beta\left(\sum_i \mathbf{1}\{P_i \le t\}\right)$. We then choose the largest threshold t such that this estimated FDP is at most α. Concretely, we define $\widehat{t} = \widehat{t}_\alpha(P)$ as

$$\widehat{t} := \max_{t \in [0,1]} \{t : \widehat{\mathrm{FDP}}(t) \le \alpha\}, \tag{3.5}$$

and we reject all hypotheses corresponding to p-values smaller than $\widehat{t}$. It is not hard to verify that $\widehat{t} \geq \frac{\alpha \cdot \widehat{k}}{n}$ (for BH) or $\widehat{t} \geq \frac{\alpha \cdot \beta(\widehat{k})}{n}$ (for BY), and that the threshold rule (3.5) corresponds to rejecting the same set of hypotheses as the original rule (3.1) or (3.2) (for the BH or BY procedures, respectively).

3.3. *A super-uniformity lemma for FDR control.* For a set x of n elements, we let $x^{-i} := \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ denote the set of $n-1$ elements that leaves out the i-th element. For a vector $x \in \mathbb{R}^n$, we use $\widetilde{x}^{-i} := (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \in \mathbb{R}^n$ to denote a vector with the i-th coordinate set to zero. Of course, $\widetilde{x}^{-i}$ and x^{-i} contain the same information and can be reconstructed from each other.

DEFINITION 4 (LOOP). A function $f : [0,1]^n \to [0,\infty)$ is said to satisfy the *leave-one-out property* (LOOP) if for any null index $i \in \mathcal{H}^0$ and any $x \in [0,1]^n$, we have

$$f(\widetilde{x}^{-i}) > 0 \text{ and } \begin{cases} \text{if } x_i \leq f(x), \text{ then } f(\widetilde{x}^{-i}) = f(x), \\ \text{if } x_i > f(x), \text{ then } x_i > f(\widetilde{x}^{-i}). \end{cases} \tag{3.6}$$

We will refer to LOOP in situations where x is the vector of p-values P, and $f(P)$ represents a threshold. When f satisfies LOOP, even though threshold $f(\widetilde{P}^{-i})$ may differ significantly from $f(P)$, the p-value P_i will either lie below both thresholds, or above both thresholds—in other words, from the perspective of P_i, the threshold might as well have been $f(\widetilde{P}^{-i})$ instead of $f(P)$.

To develop some intuition for the lemma that follows, we note that our super-uniformity assumption (2.3) on null p-values can be reformulated as:

$$\text{For any } i \in \mathcal{H}^0, \quad \mathbb{E}\left[\frac{\mathbf{1}\{P_i \leq t\}}{t}\right] \leq 1 \text{ for any fixed } t \in [0,1]. \tag{3.7}$$

Of course, if P_i is uniform then the above inequality holds with equality.

The following lemma guarantees that property (3.7) continues to hold for certain random thresholds $f(P)$. Recall that the term "nonincreasing" is interpreted coordinatewise, with respect to the orthant ordering (Def. 1).

LEMMA 1 (Super-uniformity lemma). *Let $i \in \mathcal{H}^0$ be a null hypothesis.*

(a) For any nonincreasing function $f : [0,1]^n \to [0,\infty)$, if P_i is independent of P^{-i}, then we have

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \leq f(P)\}}{f(P)} \,\middle|\, P^{-i}\right] \leq 1.$$

Furthermore, if we additionally assume that f has range $[0,1]$ and satisfies LOOP (3.6), and that P_i is uniformly distributed, then the inequality is replaced with equality:

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le f(P)\}}{f(P)} \;\middle|\; P^{-i}\right] = 1.$$

(b) For any nonincreasing function $f : [0,1]^n \to [0,\infty)$, if P is PRDS with respect to P_i, then

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le f(P)\}}{f(P)}\right] \le 1.$$

(c) For any constant $c \ge 0$, any function $f : [0,1]^n \to [0,\infty)$, and any reshaping function β, under arbitrary dependence of the p-values,

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le c \cdot \beta(f(P))\}}{c \cdot f(P)}\right] \le 1.$$

(d) For any constant $c \ge 0$, any functions $f_1, \dots, f_m : [0,1]^n \to [0,\infty)$, and any reshaping functions $\beta_1, \dots, \beta_m$, under arbitrary dependence of the p-values,

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le c \cdot \prod_{\ell=1}^m \beta_\ell(f_\ell(P))\}}{c \cdot \prod_{\ell=1}^m f_\ell(P)}\right] \le 1.$$

The proofs of statement (a) with equality, and of statement (d), are given in Appendix A. Statement (a) with inequality is recovered as a special case of statement (b), which was proved by Blanchard and Roquain [9], who also proved (c). The more general statement (d), with more than one reshaping function present in the bound, will be required in the proof of a novel group super-uniformity Lemma 3.

This super-uniformity lemma is central to succinct proofs of the results of this paper. It can also be extended to allow for multiple reshaping functions to be used simultaneously—see Corollary **??** below. To give a first hint of the power of this lemma, we begin by providing a short proof that the BH procedure controls FDR.

PROOF OF PROPOSITION 1. We prove statement (b) first. By the definition of FDR, we have

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} \mathbf{1}\left\{P_i \le \frac{\alpha\cdot\widehat{k}}{n}\right\}}{\widehat{k}}\right] = \sum_{i\in\mathcal{H}^0} \frac{\alpha}{n} \cdot \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le \frac{\alpha\cdot\widehat{k}}{n}\right\}}{\frac{\alpha\cdot\widehat{k}}{n}}\right] \overset{(i)}{\le} \alpha\frac{|\mathcal{H}^0|}{n},$$

where step (i) follows by applying Lemma 1(b) with $f(P) = \frac{\alpha \cdot \widehat{k}(P)}{n}$. Turning to statement (c), we repeat the same argument, but now incorporating the reshaping function β:

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} \mathbf{1}\left\{P_i \le \frac{\alpha\cdot\beta(\widehat{k})}{n}\right\}}{\widehat{k}}\right] = \sum_{i\in\mathcal{H}^0} \frac{\alpha}{n} \cdot \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le \frac{\alpha\cdot\beta(\widehat{k})}{n}\right\}}{\frac{\alpha\cdot\widehat{k}}{n}}\right] \overset{(ii)}{\le} \alpha\frac{|\mathcal{H}^0|}{n},$$

where step (ii) follows from Lemma 1(c) with $c = \alpha/n$, $f_1(P) = \widehat{k}(P)$ and $\beta_1 = \beta$.

In order to prove statement (a) with equality by applying Lemma 1(a), it is straightforward to see that $f(P)$ satisfies the LOOP condition (3.6) (for completeness, this statement is proved in a more general setting in Proposition 2). □

We note that leave-one-out style arguments, where FDR control over the entire set of rejected hypotheses $\widehat{\mathcal{S}}$ is proved by considering each P_i individually, are not new (see, for example, Sarkar [30]); however, most past works have either considered specific functions f, or have proved an inequality in their version of statement (a). Our proof is a generalization of arguments made in the special case of the BH procedure by Heesen and Janssen [18].

4. Using both prior and penalty weights. We now consider a setting where there are weights associated with each hypothesis. Prior work has focused on two natural interpretations for such weights—as priors on the nullity of hypotheses, or as penalties for wrongly rejecting nulls. Accordingly, we introduce a sequence of positive penalty weights $\{u_i\}_{i=1}^n$ and positive prior weights $\{w_i\}_{i=1}^n$, along with the normalization conditions $\sum_{i=1}^n u_i = n$ and $\sum_{i=1}^n w_i = n$. These two weight sequences can be interpreted as follows:

(U). The penalty weights u_i encode unequal importances of hypotheses: hypotheses with $u_i > 1$ indicate that they are of more interest to the scientist than hypotheses with $u_i < 1$, and hence we count these differently when measuring number of (false or total) discoveries.

(W). The prior weights w_i encode prior evidence against the null: a large $w_i > 1$ implies that a prior belief that the i-th hypothesis is more likely non-null, and a small $w_i < 1$ indicates the opposite.

Such weights might be learned from data collected in an older study on different subjects. We treat the weights as constants, but it is straightforward to extend our results to a Bayesian setting in which both $\mathcal{H}^0$ and the weights are random, but the null p-values are distributed (super)uniformly even after conditioning on these weights.

In order to model the setting (U), we define a modified criterion, following Benjamini and Hochberg [4], called the *weighted FDR* :

$$\mathrm{FDR}_u := \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} u_i \mathbf{1}\left\{i \in \widehat{\mathcal{S}}\right\}}{\sum_i u_i \mathbf{1}\left\{i \in \widehat{\mathcal{S}}\right\}}\right]. \tag{4.1}$$

In order to model setting (W), on the other hand, one does not need to alter the definition of the FDR, since these weights simply allow us to re-prioritize which hypotheses to reject, but does not alter our measure of error. We now present a doubly-weighted version of the BH and BY procedures, that can incorporate both sets of weights simultaneously.

4.1. *Doubly-weighted BH and BY procedures.* Following Blanchard and Roquain [9], we define the prior+penalty weighted BH procedure (BH_{uw}) as follows. We set $Q_i := P_i/w_i$, and then reject all hypotheses having $Q_i \le \widehat{t}$, where the threshold $\widehat{t} = \widehat{t}_{\alpha,u,w}(P)$ is given by

$$\widehat{t} := \max_{t\in[0,1]}\left\{t : \widehat{\mathrm{FDP}}(t) \le \alpha\right\}, \text{ where } \widehat{\mathrm{FDP}}(t) := \frac{t\cdot n}{\sum_i u_i \mathbf{1}(Q_i \le t)}. \tag{4.2a}$$

It is useful to rewrite this procedure as follows. Let $u_{(i)}$ represent the weights corresponding to the order statistics $Q_{(i)}$ of the weighted p-values $Q_1, \ldots, Q_n$, and define the sum of the first k weights to be $U_{(k)} := \sum_{i=1}^{k} u_{(i)}$. Then define

$$\widehat{k} = \widehat{k}_{\alpha,u,w}(P) := \max\left\{k : Q_{(k)} \le \frac{\alpha\cdot U_{(k)}}{n}\right\} \tag{4.2b}$$

and reject those hypotheses corresponding to the $\widehat{k}$ smallest weighted p-values, $\widehat{\mathcal{S}} = \left\{i : Q_i \le \frac{\alpha\cdot U_{(\widehat{k})}}{n}\right\}$. Then, the BH_{uw} procedure can be equivalently described by rejecting the first $\widehat{k}$ p-values, corresponding to a total penalty weight of $U_{(\widehat{k})}$. If we wish to use a reshaping function β to guard against arbitrary dependence among the p-values, then we can instead use a doubly weighted BY procedure, denoted as BY_{uw}, where we instead define

$$\widehat{t} := \max_{t\in[0,1]}\left\{t : \widehat{\mathrm{FDP}}(t) \le \alpha\right\} \text{ where } \widehat{\mathrm{FDP}}(t) := \frac{t\cdot n}{\beta\big(\sum_i u_i \mathbf{1}(Q_i \le t)\big)}, \tag{4.3a}$$

which corresponds to

$$\widehat{k} = \widehat{k}_{\alpha,u,w}(P) := \max\left\{k : Q_{(k)} \le \frac{\alpha\cdot\beta(U_{(k)})}{n}\right\}. \tag{4.3b}$$

(Implicitly, the choice of reshaping function β is considered to be fixed, and is suppressed in our notation for the procedure.)

With this setup, we have the following guarantees on the weighted FDR (4.1):

PROPOSITION 2. *The doubly-weighted BH and BY procedures, given in* (4.2b) *and* (4.3b)*, have the following properties:*

(a) Under independence and uniformity, if $\max_i w_i \le 1/\alpha$*, the* BH_{uw} *procedure yields* $\mathrm{FDR}_u = \frac{\alpha}{n} \sum_{i \in \mathcal{H}^0} u_i w_i$.

(b) Under positive dependence, the BH_{uw} *procedure yields* $\mathrm{FDR}_u \le \frac{\alpha}{n} \sum_{i \in \mathcal{H}^0} u_i w_i$.

(c) Under arbitrary dependence, the BY_{uw} *procedure with any reshaping function* β *yields* $\mathrm{FDR}_u \le \frac{\alpha}{n} \sum_{i \in \mathcal{H}^0} u_i w_i$.

Before we present the proof, let us discuss some special cases.

- When both sets of weights are used, Blanchard and Roquain [9] proved Proposition 2(b,c) and also Proposition 2(a) with inequality. Our proof of Proposition 2(b,c) mimics theirs, and is presented in our notation for completeness.
- When only penalty weights are used, and p-values are independent, one may set $w_i = 1$ for all i in the BH_{uw} procedure (4.2b) to recover the penalty-weighted BH procedure proposed by Benjamini and Hochberg [4] who proved Proposition 2(a), albeit with an inequality.
- When only prior weights are used, Genovese et al. [16] proposed a prior-weighted BH procedure that controls the unweighted FDR under independence. Their result can be recovered by setting $u_i = 1$ for all i in the BH_{uw} procedure (4.2b), and invoking Proposition 2(a).
- When all the weights equal one, the BH_{uw} and BY_{uw} reduce to the unweighted BH and BY procedures, and Proposition 2 then reduces to Proposition 1.

PROOF OF PROPOSITION 2. We first prove statements (b) and (c). For conciseness, to unite the proofs of the two statements, define a function γ to be the identity if we are in the setting of (b), or $\gamma = \beta$ (the reshaping function) if we are in the setting of (c). Note that we can write the weighted

FDR from equation (4.1) as

$$\mathrm{FDR}_u = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} u_i \mathbf{1}\left\{Q_i \le \frac{\alpha\cdot\gamma(U_{(\widehat{k})})}{n}\right\}}{U_{(\widehat{k})}}\right]$$
$$= \sum_{i\in\mathcal{H}^0} u_i \cdot \mathbb{E}\left[\frac{\mathbf{1}\left\{Q_i \le \frac{\alpha\cdot\gamma(U_{(\widehat{k})})}{n}\right\}}{U_{(\widehat{k})}}\right].$$

Recalling that $Q_i = P_i/w_i$ by definition, we multiply and divide by $\alpha w_i/n$ so as to obtain

$$\mathrm{FDR}_u = \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n} \cdot \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le w_i\frac{\alpha\cdot\gamma(U_{(\widehat{k})})}{n}\right\}}{w_i\frac{\alpha\cdot U_{(\widehat{k})}}{n}}\right] \overset{(i)}{\le} \alpha\frac{\sum_{i\in\mathcal{H}^0} u_i w_i}{n}. \tag{4.4}$$

Here inequality (i) follows by the observation that the function $P \mapsto U_{(\widehat{k}(P))}$ is nonincreasing in P, and then for each null hypothesis $i \in \mathcal{H}^0$, we can apply Lemma 1(b) with $f(P) = w_i\alpha U_{(\widehat{k}(P))}/n$, or Lemma 1(c) with $f_1(P) = U_{(\widehat{k}(P))}$, $\beta_1 = \beta$, and $c = w_i\alpha/n$.

Next we turn to the proof of statement (a), for which we begin by noting that if $\max_i w_i \le 1/\alpha$, then the range of f is $[0,1]$. Next, we show that $f(P) := w_i\frac{\alpha\cdot U_{(\widehat{k})}}{n}$ satisfies the LOOP condition (3.6) with respect to index i; if this holds, then under independence and uniformity, inequality (i) holds with equality by Lemma 1(a). To prove LOOP, first, it is easy to observe that $f(\widetilde{P}^{-i}) > 0$ since $\widehat{k}(\widetilde{P}^{-i}) > 0$ and u_i, w_i are strictly positive. Next, it is also straightforward that if hypothesis i was already rejected by the procedure, then setting P_i to zero does not change $\widehat{k}$. In order to apply Lemma 1, the last condition we need to verify is that assuming $Q_i > \alpha\cdot\frac{U_{(\widehat{k}(P))}}{n}$ holds (that is, if Q_i is not rejected), then $Q_i > \alpha\cdot\frac{U_{(\widehat{k}(\widetilde{P}^{-i}))}}{n}$ also holds. Indeed, if Q_i has rank a within the order statistics of Q, then we are guaranteed that

$$Q_i = Q_{(a)} > \alpha\frac{U_{(a)}}{n} \quad \text{and} \quad Q_{(b)} > \alpha\frac{U_{(b)}}{n} \text{ for all } b > a, \tag{4.5}$$

since Q_i is not rejected. Then, on setting P_i to 0, the number of rejections $\widehat{k}(\widetilde{P}^{-i})$ must be bounded by a, since the other p-values are unchanged and $Q_{(b)} > \alpha\frac{U_{(b)}}{n}$ for all $b > a$ continues to hold by condition (4.5). Then $U_{(\widehat{k}(\widetilde{P}^{-i}))} \le U_{(n)}$, and so we may then infer from condition (4.5) that $Q_i > \alpha\frac{U_{(\widehat{k}(\widetilde{P}^{-i}))}}{n}$ will also hold since the right-hand side is at most $\alpha\frac{U_{(a)}}{n}$. □

5. Null proportion adaptivity with prior and/or penalty weights. Recall that the original unweighted BH procedure provided FDR control at level $\alpha|\mathcal{H}^0|/n$, which can be significantly smaller than the target α when $|\mathcal{H}^0| \ll n$, leading to a possible loss in power compared to a procedure that fully utilized its FDR budget of α. When the p-values are known to be independent, Storey and collaborators [37, 38] proposed a simple method to estimate the number of nulls. Incorporating such an estimate into the BH procedure leads to a procedure that has possibly higher power than BH. We refer to this method as *Storey's adaptive BH procedure* (St-BH). For the rest of this section, we assume that the p-values are independent, and we demonstrate how to combine the benefits of adaptivity with prior and penalty weighting by analyzing a novel doubly-weighted St-BH procedure.

5.1. *Storey's adaptive BH method.* Fix a user-defined constant $\lambda \in [\alpha, 1)$, and define:

$$\widehat{\pi}_0 = \widehat{\pi}_0(P) := \frac{1 + \sum_j \mathbf{1}\{P_j > \lambda\}}{n(1-\lambda)}. \tag{5.1}$$

The quantity $\widehat{\pi}_0$ is a conservative estimate of the proportion of nulls, since for any λ, we have $\mathbb{E}[\widehat{\pi}_0] \geq \frac{|\mathcal{H}^0|}{n}$. In analogy with the BH procedure (3.4), the St-BH procedure chooses

$$\widehat{t}_\alpha(P) := \max_{t \in [0,1]} \{t : \widehat{\mathrm{FDP}}(t) \leq \alpha\}, \quad \text{where } \widehat{\mathrm{FDP}}(t) := \frac{\widehat{\pi}_0 n t}{\sum_i \mathbf{1}\{P_i \leq t\}}, \tag{5.2}$$

and then rejects all p-values smaller than $\widehat{t}_\alpha(P)$. This procedure comes with the following guarantee.

PROPOSITION 3. *Under independence, the St-BH procedure* (5.2) *guarantees that* $\mathrm{FDR} \leq \alpha$.

Storey and coauthors [37, 38] establish the above result using martingale arguments, and Blanchard and Roquain [10] gave an alternate proof using Lemma 1. We do not prove this proposition here, since it follows as a special case of Theorem 5.1 below. To proceed with a discussion of this result, we need the following notation. Define the leave-one-out estimate of the null-proportion as

$$\widehat{\pi}_0^{-i} = \widehat{\pi}_0(\widetilde{P}^{-i}) := \frac{1 + \sum_{j \neq i} \mathbf{1}\{P_j > \lambda\}}{n(1-\lambda)},$$

which is obtained if the i-th p-value is set to zero, and then $\widehat{\pi}$ is then estimated as in (5.1).

For future reference, we note that using simple properties of the binomials, it is easy to show that the estimate $\widehat{\pi}_0^{-i}$ satisfies the bound

$$\mathbb{E}\left[\frac{1}{\widehat{\pi}_0^{-i}}\right] = \mathbb{E}\left[\frac{1}{\widehat{\pi}_0^{-i}}\right] \leq \frac{n}{|\mathcal{H}^0|}. \tag{5.3}$$

It turns out that the bound (5.3) is the only property of the estimator $\widehat{\pi}_0$ that we require to prove FDR control. (In fact, the bound (5.3) is satisfied by several other estimators of π_0 as well—for examples, see Sarkar [30]. Blanchard and Roquain [10] propose and empirically compare several such estimates of π_0, to find that Storey's method had the best power amongst all considered alternatives.)

5.2. *A doubly-weighted adaptive BH procedure.* In this section, we demonstrate how to simultaneously incorporate both prior weights w and penalty weights u into Storey's adaptive estimator. We introduce the shorthand $|u \cdot w|_\infty := \max_j u_j w_j$, and define the St-BH$_{uw}$ procedure as follows:

1. Define the estimated null proportion as
$$\widehat{\pi}_0 := \frac{|u \cdot w|_\infty + \sum_j u_j w_j \mathbf{1}\{P_j > \lambda\}}{n(1-\lambda)}$$
2. Estimate the false discovery proportion by
$$\widehat{\mathrm{FDP}}(t) := \frac{\widehat{\pi}_0 \cdot n \cdot t}{\sum_i u_i \mathbf{1}\{P_i \leq \min\{w_i \cdot t, \lambda\}\}}$$
3. Choose the threshold for rejection as
$$\widehat{t} = \widehat{t}_{\alpha,u,w}(P) := \max\{t \in [0,\lambda] : \widehat{\mathrm{FDP}}(t) \leq \alpha\}; \tag{5.4}$$
4. Reject all hypotheses such that $P_i \leq \min\{w_i \widehat{t}, \lambda\}$.

In words, we apply the prior+penalty weighted BH$_{uw}$ procedure, while also incorporating a weighted null proportion estimator.

The St-BH$_{uw}$ procedure has the following guarantee:

THEOREM 5.1. *Under independence, the* St-BH$_{uw}$ *procedure* (5.4) *controls* FDR$_u$ *at level* α.

Notice that if we set the weights to unity, we recover exactly the St-BH procedure, and Theorem 5.1 reduces to Proposition 3. Also, if we set $\widehat{\pi}_0 = \lambda = 1$, we recover exactly the BH_{uw} procedure. To prepare for the proof of this theorem, we define the analogous leave-one-out estimate of the weighted null proportion as

$$\widehat{\pi}_0^{-i} := \frac{|u \cdot w|_\infty + \sum_{j \neq i} u_j w_j \mathbf{1}\{P_j > \lambda\}}{n(1-\lambda)}.$$

As before, $\widehat{\pi}_0^{-i}$ is a function of only P^{-i}, and this observation is crucial in the proof of Theorem 5.1. The other important property required by the proof is the following inequality:

$$\mathbb{E}\left[\frac{1}{\widehat{\pi}_0^{-i}}\right] \leq \frac{n}{\sum_{j \in \mathcal{H}^0} u_j w_j}. \tag{5.5}$$

To establish this, we need a more sophisticated result about weighted binomials, which we now present. (The bound (5.3) can be recovered as a special case of this lemma, by setting all weights to one.)

LEMMA 2 (Inverse binomial lemma). *Given a vector $a \in [0,1]^m$, constant $b \in [0,1]$, and Bernoulli variables $Z_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli(b), the weighted sum $Z := 1 + \sum_{i=1}^m a_i Z_i$ satisfies*

$$\frac{1}{1 + b\sum_{i=1}^m a_i} \leq \mathbb{E}\left[\frac{1}{Z}\right] \leq \frac{1}{b(1+\sum_{i=1}^m a_i)} \tag{5.6}$$

Since $\mathbb{E}[Z] = 1 + b\sum_{i=1}^m a_i$, the lower bound on $\mathbb{E}[1/Z]$ follows by Jensen's inequality. We include this bound to provide context for the upper bound on $\mathbb{E}[1/Z]$. The detailed proof can be found in Appendix B.

In order to see that required property (5.5) follows from Lemma 2, define

$$Z := 1 + \sum_{j \in \mathcal{H}^0, j \neq i} a_j \mathbf{1}\{P_j > \lambda\} \text{ with } a_j = \frac{u_j w_j}{|u \cdot w|_\infty},\ b = (1-\lambda),\ m = |\mathcal{H}^0| - 1.$$

Since $Z \leq \frac{n(1-\lambda)}{|u \cdot w|_\infty}\widehat{\pi}_0^{-i}$, applying Lemma 2 guarantees that

$$\mathbb{E}\left[\frac{|u \cdot w|_\infty}{n(1-\lambda)\widehat{\pi}_0^{-i}}\right] \leq \mathbb{E}\left[\frac{1}{Z}\right] \leq \frac{|u \cdot w|_\infty}{(1-\lambda)(|u \cdot w|_\infty + \sum_{j \in \mathcal{H}^0, j \neq i} u_j w_j)}.$$

Some simple algebra then leads to property (5.5). We are now ready to prove Theorem 5.1.

PROOF OF THEOREM 5.1. By the definition of FDR_u, we have

$$\begin{aligned}\mathrm{FDR}_u &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}^0} u_i \mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}{\sum_i u_i \mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}\right] \\ &\overset{(i)}{\le} \sum_{i\in\mathcal{H}^0} \mathbb{E}\left[\frac{\alpha u_i \mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}{\widehat{\pi}_0\cdot\widehat{t}\cdot n}\right] \\ &\overset{(ii)}{=} \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n}\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}{\widehat{\pi}_0\cdot w_i\cdot\widehat{t}}\right],\end{aligned}$$

where step (i) uses the fact that $\frac{\widehat{\pi}_0\cdot n\cdot\widehat{t}}{\sum_i u_i \mathbf{1}\{P_i\le\min\{w_i\cdot\widehat{t},\lambda\}\}} \le \alpha$, by construction of the procedure; and step (ii) follows by multiplying and dividing by w_i in term i. Moreover, since $\widehat{\pi}_0^{-i} = \widehat{\pi}_0$ whenever the numerator is one, we have

$$\begin{aligned}\mathrm{FDR}_u &\le \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n}\cdot\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}{\widehat{\pi}_0^{-i}\cdot w_i\cdot\widehat{t}}\right] \\ &\overset{(iii)}{=} \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n}\cdot\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le \min\{w_i\cdot\widehat{t},\lambda\}\right\}}{w_i\cdot\widehat{t}}\middle| P^{-i}\right]\frac{1}{\widehat{\pi}_0^{-i}}\right] \\ &\overset{(iv)}{\le} \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n}\cdot\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le w_i\cdot\widehat{t}\right\}}{w_i\cdot\widehat{t}}\middle| P^{-i}\right]\frac{1}{\widehat{\pi}_0^{-i}}\right].\end{aligned}$$

where step (iii) follows since conditioning on P^{-i} fully determines $\widehat{\pi}_0^{-i}$, and step (iv) follows from the elementary bound $\min\{w_i\cdot\widehat{t},\lambda\} \le w_i\cdot\widehat{t}$. Noting that for each P^{-i}, the function $P_i \mapsto \widehat{t}(P_i, P^{-i})$ is nonincreasing in P_i, applying Lemma 1 ensures that $\mathbb{E}\left[\frac{\mathbf{1}\{P_i\le w_i\cdot\widehat{t}\}}{w_i\cdot\widehat{t}}\middle| P^{-i}\right] \le 1$, and hence that

$$\begin{aligned}\mathrm{FDR}_u &\le \sum_{i\in\mathcal{H}^0} \frac{\alpha u_i w_i}{n}\cdot\mathbb{E}\left[\frac{1}{\widehat{\pi}_0^{-i}}\right] \\ &\overset{(v)}{\le} \alpha,\end{aligned}$$

where inequality (v) follows from property (5.5), concluding the proof.

□

6. Combining one layer of groups, with prior and/or penalty weights and null proportion adaptivity. We now demonstrate how to incorporate prior structural knowledge, where the hypotheses are separated into a partition of (non-overlapping) known groups, into the procedures introduced so far. In the next section, we will generalize some of these ideas to handle multiple arbitrary possibly-incomplete partitions consisting of possibly-overlapping groups.

When provided with group-level p-values (meaning a p-value corresponding to a test of whether the group is entirely null), one may just trivially run a BHY procedure on these p-values to control the group FDR. This section will tackle the setting where the group-level p-values are formed by combining individual-level p-values using the Simes procedure, and we would like to control the group-level FDR, possibly in the presence of prior and/or penalty weights. Our main technical tool will be a group-level super-uniformity lemma (Lemma 3), in analogy with the elementwise super-uniformity lemma (Lemma 1).

6.1. *Formal Setup.* Suppose that we have partitioned our hypotheses into G groups of size $n_1, n_2, \ldots, n_G$, with $n = n_1 + \cdots + n_G$ and that we have access to weights at the group level: prior weights $u_1, \ldots, u_G$, and penalty weights $w_1, \ldots, w_G$. (The Simes p-values for each group may themselves be formed using a weighted procedure, but these weights will be discussed separately later on.)

To summarize, our data takes the form

$$\underbrace{P_1, \ldots, P_{n_1}}_{\text{Group } A_1,\ \text{weights } u_1, w_1} \quad , \ldots , \quad \underbrace{P_{n_1+\cdots+n_{G-1}+1}, \ldots, P_n}_{\text{Group } A_G,\ \text{weights } u_G, w_G}, \tag{6.1}$$

and we wish to select a subset of these groups, $\widehat{\mathcal{S}}_{\text{grp}} \subseteq [G]$, so that the proportion of null groups (groups consisting entirely of null hypotheses) is not too high. We define the set of null groups as

$$\mathcal{H}^0_{\text{grp}} = \left\{ g \in [G] : A_g \subseteq \mathcal{H}^0 \right\},$$

that is, any group A_g that contains only null hypotheses. The weighted group FDR is then given by

$$\text{FDR}^{\text{grp}}_u := \mathbb{E}\left[\frac{\sum_{g \in \mathcal{H}^0_{\text{grp}}} u_g \mathbf{1}\left\{ g \in \widehat{\mathcal{S}}_{\text{grp}} \right\}}{\sum_g u_g \mathbf{1}\left\{ g \in \widehat{\mathcal{S}}_{\text{grp}} \right\}} \right].$$

In order to test each group A_g for rejection, we need to compute a p-value P^{grp}_g for this group—in order to be a true p-value, we need to ensure that

P_g^{grp} is (super)uniformly distributed whenever $g \in \mathcal{H}_{\text{grp}}^0$, that is whenever g is a null group.

Many statistics have been proposed that combine elementwise p-values [17]. When the p-values are independent, some options include Fisher's $-2\sum_i \ln P_i$ and Rosenthal's $\sum_i \Phi^{-1}(P_i)$, where Φ is the Gaussian CDF (originally proposed by Stouffer et al. [39]). When there are very few non-nulls, the Bonferroni correction is known to be more powerful, and it also works under arbitrary dependence, as does Rüschendorf's [29, 41] proposal of $2\sum_i P_i/n$, and Rüger's [28, 25] proposal of $P_{(k)} \cdot n/k$ for a fixed k.

In our setting, the group-level p-value can be arbitrary, either built by combining the individual p-values of the hypotheses within the group, or constructed from new independent data. *No matter how the group-level p-values are constructed*, one may simply apply a BHY procedure, appropriately weighted or reshaped or adapted, to the group-level p-values to control the group-level FDR. One example of special interest, closely related to the BHY procedure, is the setting where the group p-values are formed from elementwise p-values using the *Simes procedure*, which we next discuss.

6.2. *Generalized* Simes *tests for the global null.* Simes [33] proposed an improvement to the Bonferroni procedure for global null testing at level α. We first calculate the Simes p-value using a reshaping function $\widetilde{\beta}$ if required:[2]

$$\text{Simes}(P) = \min_{1 \le k \le n} \frac{P_{(k)} \cdot n}{\widetilde{\beta}(k)},$$

and we reject H_{GN} if $\text{Simes}(P) \le \alpha$. The connection to the BHY procedure is quite transparent: note that $\text{Simes}(P) \le \alpha$ if and only if the BHY procedure makes at least one rejection at level α. It is well known that the Simes p-value, $\text{Simes}(P)$, really is a bonafide p-value, a result that we will recover as a special case of Proposition 4.

6.3. *The prior-weighted* Simes_w *test for the global null.* The Simes test [33] was extended by Hochberg and Liberman [19] to incorporate prior weights under independence. As before, we define weighted p-values $Q_i := P_i/w_i^{(1)}$ for each hypothesis, and then calculate the generalized Simes_w p-value for the group as

$$\text{Simes}_w(P) := \min_{1 \le k \le n} \frac{Q_{(k)} \cdot n}{k}. \tag{6.2}$$

[2]Here and henceforth, the tilde in $\widetilde{\beta}$ will be used to signified a reshaping function for calculating a Simes p-value *within* a single group, and we will continue the use of notation β, without the tilde, when comparing these p-values *across* multiple groups.

The global null hypothesis for the group A_g, i.e. the hypothesis that $A_g \subseteq \mathcal{H}^0$ consists entirely of nulls, is then rejected at the level α if $\text{Simes}_w(P) \leq \alpha$.

In a more general setting where the individual p-values P_i within the group A_g may be arbitrarily dependent, we can instead consider the reshaped weighted Simes p-value, given by

$$\text{rSimes}_w(P) := \min_{1 \leq k \leq n} \frac{Q_{(k)} \cdot n}{\widetilde{\beta}(k)} \tag{6.3}$$

for a reshaping function $\widetilde{\beta}$ (recall Definition 3).

The following result states that the (weighted and/or reshaped) Simes p-value really is a bonafide p-value.

PROPOSITION 4. *Under the global null hypothesis, the weighted Simes p-value has the following properties:*

(a) Under independence and uniformity, if $\max_i w_i \leq 1/\alpha$, $\text{Simes}_w(P)$ *is exactly uniformly distributed.*
(b) Under positive dependence, $\text{Simes}_w(P)$ *is super-uniformly distributed.*
(c) Under arbitrary dependence, the reshaped Simes p-value $\text{rSimes}_w(P)$ *is super-uniformly distributed.*

While statement (a) was first proven by Hochberg and Liberman [19], and statement (c) under unit weights by Hommel [20], all the above statements are straightforward consequences of the properties of the weighted BH and BY procedures. For completeness, we prove this proposition below.

PROOF OF PROPOSITION 4. First we consider the Simes p-value without reshaping, in the setting of positive dependence. Examining the definition of the BH_{uw} procedure with the same prior weights w_i and with uniform penalty weights $u_i = 1$, we see that the weighted Simes p-value, Simes_w, is the minimum threshold α for which P passes the BH_w procedure:

$$\text{Simes}_w(P) = \min \{\alpha \in [0,1] : \text{BH}_w \text{ makes at least one rejection at level } \alpha\}.$$

In other words, we have the equivalence

$$\text{Simes}_w(P) \leq t \;\Leftrightarrow\; \widehat{k}_{t,u,w}(P) \geq 1 \tag{6.4}$$

for any $t \in [0,1]$, recalling from (4.2b) that $\widehat{k}_{t,u,w}$ is the number of rejections of the $\text{BH}_{u,w}$ procedure at the target FDR level t. Now, note that under the global null, if we run BH_{uw} at some level t, any discovery made by BH_{uw} is

a false discovery. Hence the FDP is equal to one whenever there is at least one discovery (and zero otherwise), and we have $\mathrm{FDR} = \Pr\left\{\widehat{k}_{t,u,w} > 0\right\}$. Our previous result, Proposition 2(b), proves that $\mathrm{FDR} \le t$ for the BH_{uw} procedure. Hence under the global null, we have

$$\Pr\{\mathrm{Simes}_w(P) \le t\} = \Pr\left\{\widehat{k}_{t,u,w} > 0\right\} \le t$$

Under independence and uniformity, Proposition 2(a) also implies that this statement holds with equality.

Next, we turn to the setting of arbitrary dependence, where we now use the reshaped weighted Simes p-value, $\mathrm{rSimes}_w(P)$. In this setting, we can similarly see that

$$\mathrm{rSimes}_w(P) = \min\left\{\alpha \in [0,1] : \mathrm{BY}_w \text{ makes at least one rejection at level } \alpha\right\},$$

where now we compare against the generalized BY procedure that uses the same reshaping function $\widetilde{\beta}$ as we use in our reshaped Simes test. Applying Proposition 2(c), we again show that $\Pr\{\mathrm{rSimes}_w(P) \le t\}$ is equal to the FDR of the corresponding BY procedure, and therefore, Proposition 2(c) implies that this quantity is $\le t$. □

We now demonstrate how to employ the weighted Simes p-value to control group FDR, first under independence between groups for clarity, and later under more general assumptions. For this purpose, the following group-level super-uniformity lemma will prove useful.

6.4. *A group-level super-uniformity lemma.* In analogy to the super-uniformity Lemma 1, we present the following lemma, which contains analogous bounds under the settings of independent or positively dependent base p-values (in which case the group p-value is constructed with a Simes p-value), and in the setting of arbitrarily dependent base p-values (in which case the group p-value can be constructed by any method—Simes, Fisher, or others—as long as it is a valid p-value.)

LEMMA 3 (Group-level super-uniformity lemma). *Let $g \in \mathcal{H}^0_{\mathrm{grp}}$ be a null group, that is, $A_g \subseteq \mathcal{H}^0$. Let P_{A_g} denote the p-values in this group, $P_{A_g} = (P_j)_{j \in A_g}$, and let P_{-A_g} denote the remaining p-values, $P_{-A_g} = (P_j)_{j \notin A_g}$.*

(a) If $f : [0,1]^n \to [0,\infty)$ is a nonincreasing function, and the base p-values $P_1, \ldots, P_n$ are independent, then

$$\mathbb{E}\left[\frac{\mathbf{1}\left\{\mathrm{Simes}_w(P_{A_g}) \le f(P)\right\}}{f(P)} \;\middle|\; P_{-A_g}\right] \le 1.$$

(b) If $f : [0,1]^n \to [0,\infty)$ is a nonincreasing function, and the base p-values $P_1, \dots, P_n$ are positively dependent, then

$$\mathbb{E}\left[\frac{\mathbf{1}\left\{\text{Simes}_w(P_{A_g}) \le f(P)\right\}}{f(P)}\right] \le 1.$$

(c) If the base p-values $P_1, \dots, P_n$ are arbitrarily dependent, then for any constant $c > 0$, any reshaping function β, and any function $f : [0,1]^n \to [0,\infty)$, we have

$$\mathbb{E}\left[\frac{\mathbf{1}\left\{T(P_{A_g}) \le c\beta(f(P))\right\}}{cf(P)}\right] \le 1,$$

where $T : [0,1]^{|A_g|} \to [0,1]$ is any valid group p-value, i.e. any function with the property that $T(P_{A_g})$ is superuniform whenever g is a null group. (For example, we may take $T(P_{A_g}) = \text{rSimes}_w(P_{A_g})$, the reshaped weighted Simes p-value formed using reshaping function $\widetilde{\beta}$.)

(d) Let $g_1, \dots, g_k$ be a set of k possibly overlapping null groups, meaning $A_{g_1}, \dots, A_{g_k} \subseteq \mathcal{H}^0$, and $S_1, \dots, S_k$ represent the corresponding Simes' p-values. If $f : [0,1]^n \to [0,\infty)$ is a nonincreasing function, and the base p-values $P_1, \dots, P_n$ are positively dependent, then

$$\mathbb{E}\left[\frac{\mathbf{1}\left\{\text{Simes}(S_1, \dots, S_k) \le f(P)\right\}}{f(P)}\right] \le 1.$$

The proof of this lemma relies on Lemma 1, and can be found in Appendix C. We remark that statement (d) is different from statement (b) applied to the null group $g = \bigcup_{i=1}^k g_i$; indeed, in statement (d), the arguments to the Simes' procedure are themselves Simes' p-values, and not the original base p-values. If desired, statement (d) can be further bootstrapped to apply to the root of an entire tree of null groups, where each internal node stores the Simes' p-value calculated on its children.

6.5. *Group FDR control for a single partition.* With the tools developed so far in place, it is now simple to construct a procedure that controls FDR at the group level, given a single partition of our n hypotheses into groups $A_1, \dots, A_G$: we first form some summary p-value P_g^{grp} for each group $g = 1, \dots, G$, and then run either a BH procedure or a BY procedure on the resulting list $P_1^{\text{grp}}, \dots, P_G^{\text{grp}}$, according to the dependence assumptions.

First we construct the group p-values. For each group, we may choose to use a weighted Simes p-value to collapse the group into a single summary p-value, reducing the overall problem to a standard multiple testing

problem on G p-values. To avoid conflicting notation with the group-level weights $w_1, \ldots, w_G$, let $w_1^{(0)}, \ldots, w_n^{(0)}$ denote the weights on the n individual hypotheses for computing these group-level Simes p-values.

Referring to the notation of setup (6.1), define $Q_i := P_i / w_i^{(0)}$ and compute the Simes_w p-values for each group, $\text{Simes}_w(P_{A_1}), \ldots, \text{Simes}_w(P_{A_G})$, where

$$P_g^{\text{grp}} = \text{Simes}_w(P_{A_g}) = \min_{1 \le k \le |A_g|} \frac{Q_{g,(k)} \cdot \sum_{j \in A_g} w_j^{(0)}}{k}, \tag{6.5}$$

where $Q_{g,(k)}$ is the k-th smallest p-value in group A_g. If we choose to use a reshaping function $\widetilde{\beta}$, then the weighted rSimes p-value is computed as

$$P_g^{\text{grp}} = \text{rSimes}_w(P_{A_g}) = \min_{1 \le k \le |A_g|} \frac{Q_{g,(k)} \cdot \sum_{j \in A_g} w_j^{(0)}}{\widetilde{\beta}_g(k)} \tag{6.6}$$

for each group, where the subscript g in the reshaping function, i.e. $\widetilde{\beta}_g(k)$, indicates that for each group we are free to choose different $\widetilde{\beta}$ functions (though there may be little reason to exercise this freedom).

Note that, in both definitions, the numerator uses $\sum_{j \in A_g} w_j^{(0)}$ instead of $|A_g|$ because, depending on how we normalize our weights $w_i^{(0)}$, we may have that the sum of all weights over all groups is n, but the sum of the weights within the group A_g need not be $|A_g|$; hence the adjustment with the appropriate normalization. As argued earlier, $\text{Simes}_w(P_{A_g})$ is a bonafide p-value, i.e. is superuniform, under the "group intersection hypothesis" that A_g is composed only of nulls, whenever p-values in A_g are independent or positively dependent; under arbitrary dependence, the reshaped p-value $\text{rSimes}_w(P_{A_g})$ is now the one that is guaranteed to be superuniform.

Finally, if we do not wish to use a Simes p-value, we are free to set

$$P_g^{\text{grp}} = T(P_{A_g})$$

for any function $T : [0,1]^{|A_g|} \to [0,1]$ with the property that it is a valid p-value under the group null, that is,

$$g \in \mathcal{H}^0_{\text{grp}} \;\Rightarrow\; T(P_{A_g}) \text{ is superuniform.}$$

This includes the reshaped Simes p-value as a special case, i.e. $T(P_{A_g}) = \text{rSimes}_w(P_{A_g})$, and also encompasses the Fisher and Rosenthal group p-value constructions. (In principle, the group-level p-values could also be computed from the raw data, which does not alter any of the results proved here, as long as P_g^{grp} is still superuniform for any null group $g \in \mathcal{H}^0_{\text{grp}}$).)

Next, we state the results for FDR control in this group testing setting. As expected, under independence we are free to use adaptivity, while under arbitrary dependence, we are obligated to use reshaping. Under independence, the null proportion estimate (for fixed $\lambda_{\mathrm{grp}} \in (0,1)$) is naturally defined as

$$\widehat{\pi}_{\mathrm{grp}} := \frac{|u \cdot w|_\infty + \sum_{g=1}^{G} u_g w_g \mathbf{1}\left\{P_g^{grp} > \lambda_{\mathrm{grp}}\right\}}{G(1-\lambda_{\mathrm{grp}})}. \tag{6.7}$$

PROPOSITION 5 (Group-level FDR control). *Recalling setup* (6.1)*:*
(a) Under independence of the group p-values, the St-BH$_{uw}$ *procedure applied to* $\{P_g^{grp}\}$ *achieves* $\mathrm{FDR}_u^{grp} \le \alpha$.
(b) Under positive dependence of the base p-values in P*, the* BH$_{uw}$ *procedure applied to Simes' p-values* $\{\mathrm{Simes}(P_{A_g})\}$ *achieves* $\mathrm{FDR}_u^{grp} \le \alpha$.
(c) Under arbitrary dependence of group p-values, the BY$_{uw}$ *procedure applied to* $\{P_g^{grp}\}$ *achieves* $\mathrm{FDR}_u^{grp} \le \alpha$.

We do not provide an explicit proof of this proposition, since the proofs for $\mathrm{FDR}_u^{\mathrm{grp}}$ control directly mimic the corresponding proofs of FDR_u control for the same procedures applied to the base p-values in P, with the single alteration that the appropriate statements of Lemma 3 are invoked in place of their corresponding statements in Lemma 1. Further, the proof of the above proposition follows as a special case of the proof of Theorem 7.1 for the multilayer procedure `p-filter` to be introduced in the next section.

As observed in Barber and Ramdas [1] for the unweighted case, we can view the group FDR procedure described in this subsection as an interpolation between the weighted Simes test, and the weighted St-BH procedure. This is seen by considering two extremes: taking one single group of size n recovers the Simes test of the global null (across all n hypotheses), while taking instead n groups of size one recovers the BH or BY procedures for testing individual hypotheses. Of course, while the above algorithm controls group-level FDR, it has no guarantee for the elementwise FDR when the hypotheses are considered individually rather than in groups. Our general multilayer method, introduced in the next section, gives this type of simultaneous guarantee under various forms of dependence.

7. FDR **control and internal consistency for multiple partitions.** We now turn to the general case in which we allow multiple partitions of the set of hypotheses, as in Barber and Ramdas [1], while continuing to support prior weights, penalty weights, group-level adaptivity, reshaping under arbitrary dependence, as well as other generalizations of overlapping groups and

incomplete partitions. The presented framework will specialize, in the case of a single partition, to all the procedures discussed in previous sections. Our formal setup is as follows. We have M partitions of interest, with the mth partition having G_m many groups:

$$A_1^{(m)}, \ldots, A_{G_m}^{(m)} \subseteq [n] \text{ for } m = 1, \ldots, M.$$

The partitions and the groups within each partition can be specified in any order. The order in which the partitions are described makes no difference to the final output; similarly, the ordering within each partition's groups is also arbitrary.

As in Section 6, define the null set for the mth partition as containing those groups that are entirely filled with null hypotheses:

$$\mathcal{H}_m^0 = \left\{ g \in [G_m] : A_g^{(m)} \subseteq \mathcal{H}^0 \right\}.$$

As before, there is a natural notion of group-level FDR—namely, a falsely discovered group is a group that is entirely composed of true nulls, for which our algorithm has incorrectly proclaimed at least one discovery within that group (i.e. effectively has rejected the global null hypothesis within that group). Each group in each partition is associated with a summary p-value P_g^m, which we assume to be derived from the base p-values $(P_i : i \in A_g^m)$, but may more generally be calculated from the raw data or from another source without altering our results.

For interpretability, it is arguably of interest to maintain "internal consistency" among rejected hypotheses and groups. For a single partition into non-overlapping groups, we say that a set of rejections at the individual level and group level(s) are internally consistent if:

(7.1) We only reject groups containing at least one rejected hypothesis, and we only reject a hypothesis if its group is also rejected.

We will require the rejections to be internally consistent simultaneously for all partitions, and later extend this notion in two ways to partitions with leftover sets and overlapping groups (*weak* and *strong* internal consistency). The above notion of internal consistency is an extension, to the multilayer setting, of the requirement of both *coherence* and *consonance* as defined in the classical paper by Gabriel [14], and explored in depth in the FWER literature by Sonnemann and Finner [34, 35, 36], and Romano et al. [27].

Even in the case of just one layer of groups on top of the individual hypotheses, handling consistency while controlling both group-level and individual-level FDR is not trivial. For example, a sequential procedure, of

first rejecting groups at a target FDR level α_1, and then rejecting individual hypotheses within rejected groups at level α_2, may neither succeed in controlling the FDR at the individual level (due to not accounting for selection bias), nor succeed in being internally consistent (because some groups may get selected in the first round, but none of the individuals within that group may get selected in the second round). Further, such a method is not easily generalized to non-hierarchical partitions. Similarly, a *parallel* procedure of choosing which groups to reject independently of which individuals to reject may also fail to be internally consistent. A naive solution to this issue is *intersection*—i.e., rejecting those hypotheses whose groups are rejected at every layer—but this may also fail to control FDR at both levels (see Barber and Ramdas [1] for examples of this phenomenon).

Our algorithm can guarantee FDR control at the hypothesis and group levels, that holds *simultaneously for multiple arbitrary non-hierarchical partitions* of the hypotheses, while also incorporating weights and adaptivity, all while maintaining the internal consistency of rejected hypotheses. We now introduce the framework formally.

7.1. *The* `p-filter` *framework.* The algorithmic framework presented in this section generalizes the `p-filter` framework by Barber and Ramdas [1]. Roughly speaking, the algorithm is a multivariate extension of classical step-up procedures, based on the following sequence of steps:

- Select all hypotheses in each layer whose p-values are smaller than some initial layer-specific threshold.
- Reject an elementary hypothesis if it is contained in a selected group in every layer.
- In each layer, reject a group hypothesis if it contains a rejected elementary hypothesis. Then, estimate the group-FDP in each layer.
- Lower the initial thresholds at each layer, and repeat the steps above, until the group-FDP is below the desired level for all partitions.

We use the same name "`p-filter`" for this extended algorithm, while noting six important ways in which this paper adds to the flexibility of the earlier framework.

1. **Weights.** The m-th partition is associated with two sets of G_m many positive weights, one for each group g in that partition:

$$\text{Penalties } \{u_g^{(m)}\} \text{ and priors } \{w_g^{(m)}\}, \text{ such that } \sum_{g=1}^{G_m} u_g^{(m)} w_g^{(m)} = G_m.$$

   These can be used to take differing group sizes into account.

2. **Reshaping.** If the p-values within or across layers are arbitarily dependent, we use β_m to reshape thresholds in layer m (possibly different for the M layers). If Simes p-values are used to form group level p-values, then we may place weights $w_i^{(0)}$ on the base p-values $P_1, \ldots, P_n$ for the purpose of calculating a weighted Simes p-value, $P_g^m = \text{Simes}_w(P_{A_g^m})$; if additionally the p-values within the group are arbitrarily dependent, we would then choose reshaping functions $\widetilde{\beta}_{mg}$, potentially with different functions chosen for each group, for computing $P_g^m = \text{rSimes}_w(P_{A_g^m})$.
3. **Adaptivity.** For any partition whose group-level p-values are known to be independent (i.e., independence between groups, not necessarily within groups), we can incorporate Storey's null-proportion adaptivity. Every layer m has a user-defined constant $\lambda_m \in (0,1)$, with which we define a weighted null proportion estimator for partition m:

$$\widehat{\pi}_m := \frac{|u^{(m)} \cdot w^{(m)}|_\infty + \sum_g u_g^{(m)} w_g^{(m)} \mathbf{1}\left\{P_g^m > \lambda_m\right\}}{G_m(1-\lambda_m)}. \tag{7.2}$$

   The use of null-proportion adaptivity in any *one layer* may improve the power in *all layers*, since more groups being discovered in one layer then allows more individual hypotheses to be discovered, and hence more groups to be discovered in other layers.
4. **Incomplete partitions.** We allow the partitions to be incomplete—we let the m-th partition's "leftover" set $L^{(m)} \subset [n]$ represent all elements from the m-th partition that do not belong to any group in the m-th partition:

$$L^{(m)} = [n] \backslash \bigcup_g A_g^{(m)} \quad \text{and} \quad \ell_m = |L^{(m)}|.$$

   This gives additional flexibility to the user who may not want to assign some hypotheses to any groups. It is important to note that $L^{(m)}$ is not just another group—this set is not counted when calculating the group-level FDR in layer m, meaning that discoveries within this leftover set do not alter the FDR at layer m, and hypotheses in this leftover set have no internal consistency constraints imposed by layer m.
5. **Overlapping groups.** We allow the groups in any partition to overlap. Any elementary hypothesis need not be part of just a single $g \in [G_m]$; we let $g_m(i)$ denote the set of groups to which P_i belongs:

$$g_m(i) = \{g \in [G_m] : P_i \in A_g^{(m)}\}.$$

   Two natural extensions of internal consistency are as follows.

- **Weak internal consistency**: we reject H_i if and only if in every partition, the following holds: either there is at least one rejected group containing i, or $i \in L^{(m)}$, the leftover set, meaning that i does not belong to any group at this layer.
- **Strong internal consistency**: we reject H_i if and only if in every partition, the following holds: either every group that contains i is rejected, or $i \in L^{(m)}$.

We remark these are not the only two notions of internal consistency that can be handled by our framework: any *monotone* notion of internal consistency still works, meaning that decreasing the p-values can only increase the number of rejections at all levels.

6. **Arbitrary group p-values.** The new `p-filter` algorithm no longer necessarily uses Simes p-values at the group layers. In other words, each group-level p-value at each layer can be arbitrary; it can be formed by combining the elementwise p-values in any way (or, as mentioned before, might be constructed from the raw data). Depending on how these p-values are constructed, we can appropriately use adaptivity (under independence, if Simes p-values are used), or reshaping (under dependence, or if we choose to use Fisher, Rosenthal, or other combination methods), as needed.

`p-filter` will reject some subset $\widehat{\mathcal{S}} \subseteq [n]$ of hypotheses and a subset $\widehat{\mathcal{S}}_m \subseteq [G_m]$ of groups in each partition m, with the exact choice of these rejection sets defined in the next subsection. Given $\widehat{\mathcal{S}}$ and the $\widehat{\mathcal{S}}_m$'s, we then define the weighted FDR for the mth partition as

$$\mathrm{FDR}_u^{(m)} = \mathbb{E}\left[\frac{\sum_{g \in \mathcal{H}_m^0} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}_m\right\}}{\sum_{g \in [G_m]} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}_m\right\}}\right].$$

The rejections made by `p-filter` will be internally consistent, and satisfy

$$\mathrm{FDR}_u^{(m)} \leq \alpha_m \text{ simultaneously for all } m = 1, \ldots, M.$$

To describe a situation where allowing for non-hierarchical layers, and for more than two layers, may be useful, imagine that we can write the p-values down in a rectangular grid where the p-values within a row have some interpretable meaning (say, hypotheses corresponding to the same point in space), and similarly the p-values within a column have a different meaning (say, hypotheses corresponding to the same point in time). Then in addition to controlling the overall FDR, we may also want to control the "spatial" FDR using a row partition where each group is a row, and "temporal"

FDR using a column partition where each group is a column, or even the "spatio-temporal" FDR for using a rectangular partition when each group is a block in space-time. Refer to Figure 1 for visualization, and see Barber and Ramdas [1] for further discussion, numerical simulations, as well as a neuroscience application.

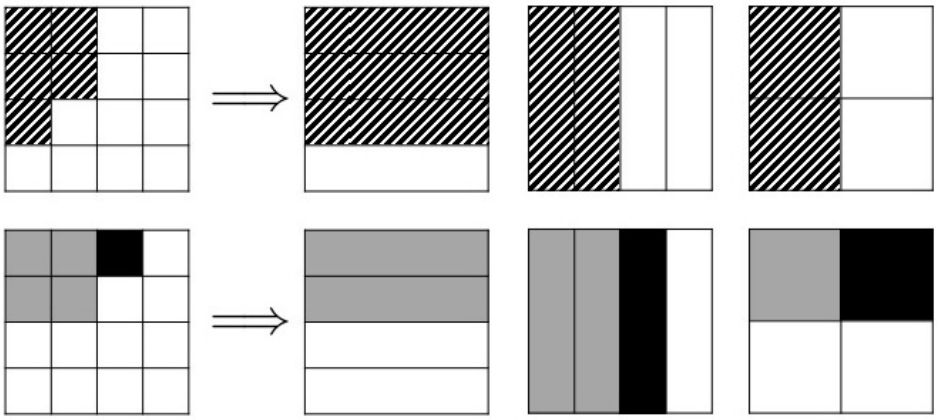

FIG 1. *Consider 16 hypotheses written as a* 4×4 *grid, with four partitions: elementary, rows, columns, blocks. On the top row is the underlying truth, with the leftmost panel showing the hypothesis-level non-nulls, and the other three panels showing which groups in each partition are hence identified as non-null. On the bottom row is an example of a set of discoveries, with the leftmost panel showing the hypothesis-level rejections, and the other three panels showing which groups are correspondingly rejected (light-grey for correct rejections, black for false rejections). The false discovery proportions in each partition are 0.2, 0, 0.33, 0.5 respectively.*

7.2. *Deriving the* `p-filter` *algorithm.* To run the `p-filter` algorithm, we need to search for rejection thresholds for each layer. These thresholds will be parametrized by *weighted* discovery counts $k_m \in [0, G_m]$ for each layer $m = 1, \ldots, M$. The reader is cautioned that each k_m will not necessarily be an integer but instead be a real number corresponding loosely to the total rejected penalty weight. If the weights $u_g^{(m)}$ are all set to equal 1, then k_m indeed corresponds to the number of groups in layer m that are rejected. Given $\vec{k} := (k_1, \ldots, k_M)$, we first perform an initial screening on each layer separately:

$$\widehat{\mathcal{S}}_m^{\text{init}}(\vec{k}) = \left\{ g \in [G_m] : P_g^m \le \min\left\{ \frac{w_g^{(m)} \alpha_m k_m}{\widehat{\pi}_m G_m}, \lambda_m \right\} \right\}. \tag{7.3a}$$

If we choose to use reshaping, we would instead define

$$\widehat{\mathcal{S}}_m^{\text{init}}(\vec{k}) = \left\{ g \in [G_m] : P_g^m \leq \min\{ \tfrac{w_g^{(m)} \alpha_m \beta_m(k_m)}{\widehat{\pi}_m G_m}, \lambda_m \} \right\}. \tag{7.3b}$$

(The null proportion estimate $\widehat{\pi}_m$ can be defined as in (7.2) if we choose to use adaptivity; if not, we simply set $\widehat{\pi}_m = 1$ and $\lambda_m = 1$, so that the same equations can be used in either setting.)

We remark that while we are not obliged to do so, one option for group-level p-values is to set $P_g^m = \text{Simes}_{w^{(1)}}(A_g^{(m)})$; for convenience, we now recompute these p-values in the notation of the present section. Recall that $Q_i = P_i / w_i^{(1)}$, and that

$$\text{Simes}_{w^{(1)}}(A_g^{(m)}) := \min_{1 \leq i \leq |A_g^{(m)}|} \frac{Q_{g,(i)}^{(m)} \sum_{j \in A_g^{(m)}} w_j^{(1)}}{i},$$

where $Q_{g,(i)}^{(m)}$ is the i-th smallest p-value in $A_g^{(m)}$. If we choose to use reshaped Simes p-values, we would instead compute

$$\text{Simes}_{w^{(1)}}(A_g^{(m)}) := \min_{1 \leq i \leq |A_g^{(m)}|} \frac{Q_{g,(i)}^{(m)} \sum_{j \in A_g^{(m)}} w_j^{(1)}}{\widetilde{\beta}_{mg}(i)},$$

where $\widetilde{\beta}_{mg}$ is the reshaping function chosen for group g in layer m.

Also recall that if the groups in the m-th layer are independent, we may use null-proportion adaptivity by setting $\widehat{\pi}_m$ as in (7.2), while otherwise we would set $\widehat{\pi}_m \equiv 1$. We then define the total set of rejections, either as

$$\widehat{\mathcal{S}}(\vec{k}) = \widehat{\mathcal{S}}_{\text{weak}}(\vec{k}) = \bigcap_{m=1}^{M} \left(\left[\bigcup_{g \in \widehat{\mathcal{S}}_m^{\text{init}}(\vec{k})} A_g^{(m)} \right] \cup L^{(m)} \right), \tag{7.4a}$$

if we wish to enforce weak internal consistency, or alternately as

$$\widehat{\mathcal{S}}(\vec{k}) = \widehat{\mathcal{S}}_{\text{strong}}(\vec{k}) = \bigcap_{m=1}^{M} \left([n] \backslash \bigcup_{g \in [G_m] \backslash \widehat{\mathcal{S}}_m^{\text{init}}(\vec{k})} A_g^m \right) \tag{7.4b}$$

for strong internal consistency. To interpret these expressions, we see that under weak consistency (7.4a), hypothesis i is rejected if, for every layer m, either i is in the leftover set $L^{(m)}$, or i belongs to at least one of the groups $A_g^{(m)}$ rejected in this layer—we can rewrite this condition as

$$\widehat{\mathcal{S}}_{\text{weak}}(\vec{k}) = \{ P_i : \forall m, \text{ either } P_i \in L^{(m)}, \text{ or } \exists\, g \in g(m,i),\, A_g^{(m)} \in \widehat{\mathcal{S}}_m^{\text{init}}(\vec{k}) \}.$$

For strong consistency (7.4b), hypothesis i is rejected if, for every layer m, either i is in the leftover set $L^{(m)}$, or i belongs to *only* rejected groups—that is, if i does not belong to any groups A_g^m which were *not* rejected—which we can rewrite as

$$\widehat{\mathcal{S}}_{\text{strong}}(\vec{k}) = \{P_i : \forall m, \text{ either } P_i \in L^{(m)}, \text{ or } \forall\, g \in g(m,i),\, A_g^{(m)} \in \widehat{\mathcal{S}}_m^{\text{init}}(\vec{k})\}.$$

Finally, we redefine the set of groups in layer m which are rejected in order to enforce internal consistency:

$$\widehat{\mathcal{S}}_m(\vec{k}) = \left\{ g \in [G_m] : A_g^{(m)} \cap \widehat{\mathcal{S}}(\vec{k}) \neq \emptyset \text{ and } g \in \widehat{\mathcal{S}}_m^{\text{init}}(\vec{k}) \right\}, \tag{7.6}$$

where either $\widehat{\mathcal{S}}(\vec{k}) = \widehat{\mathcal{S}}_{\text{weak}}(\vec{k})$ or $\widehat{\mathcal{S}}(\vec{k}) = \widehat{\mathcal{S}}_{\text{strong}}(\vec{k})$, as desired. Examining these definitions, readers may verify that the appropriate (weak or strong) notion of internal consistency (7.1) is satisfied.

Of course, these definitions depend on the initial choice of the vector $\vec{k}$. Since we would like to make a large number of discoveries, we would like to use a $\vec{k}$ that is as large as possible, while at the same time controlling the layer-specific false discovery rates,

$$\text{FDP}_u^{(m)}(\vec{k}) := \frac{\sum_{g \in \mathcal{H}_m^0} u_g^{(m)} \mathbf{1}\left\{ g \in \widehat{\mathcal{S}}_m(\vec{k}) \right\}}{\sum_{g \in [G_m]} u_g^{(m)} \mathbf{1}\left\{ g \in \widehat{\mathcal{S}}_m(\vec{k}) \right\}}.$$

Now, define the data-dependent set of feasible vectors $\vec{k} = (k_1, \ldots, k_M)$ as

$$\widehat{\mathcal{K}} = \left\{ \vec{k} \in [0, G_1] \times \cdots \times [0, G_M] : \sum_{g \in \widehat{\mathcal{S}}_m(\vec{k})} u_g^{(m)} \geq k_m \text{ for all } m \right\}, \tag{7.7}$$

where we suppress the implicit dependence of $\widehat{\mathcal{K}}$ on input parameters such as $\alpha_m, \lambda_m, \{w_g^{(m)}\}, \{u_g^{(m)}\}$. In particular, if the penalty weights are all equal to one, then the consistency condition defining the "feasible" $\vec{k}$'s is equivalent to requiring that $|\widehat{\mathcal{S}}_m(\vec{k})| \geq k_m$ for all $m = 1, \ldots, M$—i.e., the numbers of rejections in each layer at the vector $\vec{k}$ are elementwise $\geq \vec{k}$. This is analogous to the BH procedure (3.1) and the generalized BY procedure (3.2), which each find a value k such that there are at least k many rejections at the threshold $\frac{\alpha k}{n}$ (for BH) or $\frac{\alpha \beta(k)}{n}$ (for BY). The above condition can be viewed as a generalization, to the multi-partition setting, of the "self-consistency" condition described by Blanchard and Roquain [9].

Also, it is worthy of note that the `p-filter` algorithm in Barber and Ramdas [1] was derived in terms of thresholds $\vec{t}$ instead of number of rejections $\vec{k}$,

and there the corresponding feasibleity condition was that $\widehat{\mathrm{FDP}}_m(\vec{t}) \le \alpha_m$, where $\widehat{\mathrm{FDP}}_m(\vec{t})$ is an empirical-Bayes type estimate of the FDP. Indeed, if we avoid $\widehat{\pi}_m, \beta_m, w_m, u_m$ for simplicity, then associating $\widehat{t}_m$ to $\alpha_m \widehat{k}_m / G_m$ and comparing our derivation to the one in Barber and Ramdas [1], we can see that the "self-consistency" viewpoint and the "empirical-Bayes" viewpoint are equivalent and lead to the same algorithms. However, when dealing with arbitrary dependence, the proofs are vastly simpler in terms of $\vec{k}$ than in terms of $\vec{t}$, explaining the switch in choice of notation in this paper.

As with the BH and BY procedures, we then choose the largest feasible thresholds k_m, given by:

$$\widehat{k}_m = \max\left\{k_m : \exists k_1, \dots, k_{m-1}, k_{m+1}, \dots, k_M \text{ s.t. } \vec{k} \in \widehat{\mathcal{K}}\right\} \tag{7.8}$$

This choice defines our algorithm: the `p-filter` algorithm rejects the hypotheses $\widehat{\mathcal{S}}(\widehat{k}_1, \dots, \widehat{k}_M)$, as defined in (7.4a) or (7.4b), with rejections at layer m given by $\widehat{\mathcal{S}}_m(\widehat{k}_1, \dots, \widehat{k}_M)$ as defined in (7.6). Next, we summarize the theoretical guarantees of `p-filter`.

7.3. *Theoretical guarantees.* The following proposition states that the set of feasible vectors $\widehat{\mathcal{K}}$ actually has a well-defined "maximum" corner.

PROPOSITION 6. *Let the set of feasible vectors $\widehat{\mathcal{K}}$ be defined as in equation (7.7), and let the partition-specific maximum feasible vector $\widehat{k}_m$ be defined as in equation (7.8). Then we have*

$$(\widehat{k}_1, \dots, \widehat{k}_M) \in \widehat{\mathcal{K}}\,. \tag{7.9}$$

Barber and Ramdas [1] proved this result in the setting of the original `p-filter` algorithm; for completeness, we prove it in this more general setting in Appendix D.

The vector $(\widehat{k}_1, \dots, \widehat{k}_M)$ is not just feasible from the perspective of self-consistency as captured by $\widehat{\mathcal{K}}$, but it is also feasible from the perspective of FDR control. Specifically, the next theorem proves that—assuming for now that we can find $(\widehat{k}_1, \dots, \widehat{k}_M)$—selecting the set $\widehat{\mathcal{S}}(\widehat{k}_1, \dots, \widehat{k}_M)$ guarantees simultaneous control of $\mathrm{FDR}_u^{(m)}$ for all M partitions.

THEOREM 7.1. *Any procedure that finds $(\widehat{k}_1, \dots, \widehat{k}_M)$ from definition (7.8) satisfies the following properties for all $m = 1, \dots, M$ simultaneously:*

(a) If the base p-values are independent, and all group p-values are given by $P_g^m = \mathrm{Simes}_w(P_{A_g^m})$, then employing adaptivity (defining $\widehat{\pi}_m$ as in (7.2)) guarantees that $\mathrm{FDR}_u^{(m)} \le \alpha_m$.

(b) If the base p-values are +dependent and all group p-values are given by $P_g^m = \text{Simes}_w(P_{A_g^m})$*, then* $\text{FDR}_u^{(m)} \le \alpha_m \frac{\sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)}}{G_m} \le \alpha_m$.

(c) When all p-values are arbitrarily dependent, and are constructed arbitrarily (under the assumption that P_g^m *is superuniform for any null group* $g \in \mathcal{H}_m^0$*, meaning it is a valid p-value), then using reshaping as in (7.3b) guarantees that* $\text{FDR}_u^{(m)} \le \alpha_m \frac{\sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)}}{G_m} \le \alpha_m$.

(d) In the setting of part (c), if additionally the groups at layer m *are independent (that is,* $P_{A_g^m}$ *is independent from* $P_{-A_g^m}$*, for each* $g \in [G_m]$*), then using reshaping as in (7.3b) and adaptivity for layer* m *as in (7.2), guarantees that* $\text{FDR}_u^{(m)} \le \alpha_m$.

To remark on the difference between parts (c) and (d), what these two results guarantee is that if we use adaptivity for some set $\mathcal{M}_{\text{adapt}} \subset [M]$ of layers, and do not use adaptivity (i.e. set $\widehat{\pi}_m = 1$) for the remaining layers, then FDR control is maintained across *all* layers as long as, for each $m \in \mathcal{M}_{\text{adapt}}$, the layer-wise independence statement holds—$P_{A_g^m}$ is independent from $P_{-A_g^m}$, for each $g \in [G_m]$. If this condition fails for some $m \in \mathcal{M}_{\text{adapt}}$, the FDR control in other layers will in fact not be affected.

The proof, given in Appendix E, integrates the various ingredients required to handle weights, null proportions, and groups as detailed in earlier sections. The proof includes some new ideas of independent interest, specific to handling overlapping groups with dependent p-values. One application of statement (d) is when the base p-values are independent, there are no overlapping groups, and group p-values are formed using a Fisher, Rosenthal, or other combinations of the base p-values. Recently, Katsevich and Sabatti [23] prove that in case (d), the FDR is controlled even without using reshaping, albeit at a constant factor larger than the target level.

In practice, if we have accurate side information about group structures that the rejected hypotheses likely respect, then we may significantly improve our *precision*, achieving a lower FDR than the theoretical bound, without affecting our power much. However, inaccurate side information may significantly lower our power, since each p-value would have additional misguided constraints to meet. These aspects were explored in simulations by Barber and Ramdas [1].

Special cases. The setting with a single partition ($M = 1$) recovers all other algorithms discussed in this paper. When we have the *finest* partition with n groups containing one hypothesis each, `p-filter` reduces exactly to the generalized BHY procedures discussed in Section 3, which of course include BH and BY, and their prior- and/or penalty-weighted variants BHY_{uw}

described in Section 4. When we employ adaptivity, we recover the St-BH method and its weighted variant in Section 5. When we instantiate `p-filter` with the *coarsest* partitions with a single group containing all n hypotheses, we recover exactly the generalized Simes test, and its weighted variants from Section 6. When instantiated to a single partition with G groups of hypotheses, we recover exactly the test that controls group-FDR, as discussed at the end of Section 6. All the theorems and propositions in this paper are essentially deduced as special cases of Theorem 7.1.

Algorithm 1 The `p-filter` for multi-layer FDR control

Input: M possibly incomplete partitions of possibly overlapping groups of indices $[n]$;
A vector of base p-values $P \in [0,1]^n$;
Group p-values P_g^m for each group $g = 1, \dots, G_m$ in layers $m = 1, \dots, M$;
M target FDR levels $\{\alpha_m\}$;
M sets of prior weights and/or penalty weights $\{w_g^{(m)}, u_g^{(m)}\}$;
M thresholds for adaptive null proportion estimation $\{\lambda_m\}$;
M reshaping functions $\{\beta_m\}$, if desired.
Initialize: Set $k_m = G_m$, and $\widehat{\pi}_m$ as in definition (7.2).
repeat
for $m = 1, \dots, M$ **do**
Update the mth vector: defining $\widehat{\mathcal{S}}_m(\vec{k})$ as in equation (7.6) (using weak or strong consistency, as desired), let

$$k_m \leftarrow \max \left\{ k'_m \in [0, G_m] : \sum_{g \in \widehat{\mathcal{S}}_m(k_1, \dots, k_{m-1}, k'_m, k_{m+1}, \dots, k_M)} u_g^{(m)} \geq k'_m \right\}$$

end for
until the vectors $k_1, \dots, k_M$ are all unchanged for one full cycle.
Output: Adaptive vector $\widehat{k} = (k_1, \dots, k_m)$.

7.4. *An efficient implementation.* Although one can employ a brute-force grid search to find $(\widehat{k}_1, \dots, \widehat{k}_M)$, the `p-filter` algorithm presented in Algorithm 1 is able to find this vector efficiently using a coordinate-descent style procedure, and is a strict generalization of the algorithm by the same name in Barber and Ramdas [1]. Code for this procedure is publicly available at https://www.stat.uchicago.edu/∼rina/pfilter.html. The following proposition provides a correctness guarantee for Algorithm 1:

PROPOSITION 7. *The output of Algorithm 1 is the maximum feasible corner* $(\widehat{k}_1, \dots, \widehat{k}_m)$ *defined in equations* (7.8) *and* (7.9).

This result was proved by Barber and Ramdas [1] in the setting of the original `p-filter` algorithm, where the k_m's take only integer values; here, the

algorithm is slightly more subtle, with real-valued k_m's due to the presence of penalty weights $u_g^{(m)}$. The proof of the proposition for this more general setting is given in Appendix D.

8. Discussion. This paper provides simple proofs for many existing multiple testing procedure—while also generalizing the conditions under which some of them work—including the global null test of Simes [33], the weighted Simes test of Hochberg and Liberman [19], the FDR controlling procedures of Benjamini and Hochberg [3] and Benjamini and Yekutieli [7], the weighted FDR controlling procedure of Benjamini and Hochberg [4], the p-value weighting procedure of Genovese et al. [16], and the null-proportion adaptivity procedure of Storey [37]. We also described several new procedures, including prior+penalty weighted versions of BH and St-BH, and an adaptive weighted Simes test, and showed how to unify these concepts in the grouped setting. Finally, our `p-filter` algorithm unifies and generalizes all of the above procedures, building on the original `p-filter` framework introduced by Barber and Ramdas [1].

The procedures that we have analyzed and generalized do not fully cover the huge literature on FDR controlling procedures. For example, all procedures in this paper are step-up procedures, and much work has also been done on alternative styles of procedures, like step-down, step-up-down and multi-step methods. For example, Benjamini and Liu [6] propose step-down procedures that control FDR under independence. Later, procedures by Benjamini and Liu [5] and Romano and Shaikh [26] provably control FDR under arbitrary dependence, with Gavrilov et al. [15] extending them to adaptive control under independence. Two-step adaptive procedures have been analyzed in Benjamini et al. [8] under independence, and by Blanchard and Roquain [10] under dependence. Different methods of incorporating weights into such procedures have also been studied, e.g. a different notion of the weighted Simes p-value proposed by Benjamini and Hochberg [4]. Hu et al. [21] and Benjamini and Bogomolov [2] also propose ways to take a single partition of groups into account, while Yekutieli [42] discusses hierarchical testing. The super-uniformity lemmas (Lemma 1 and, in the grouped setting, Lemma 3), can be used to quickly prove FDR control for many of these procedures, and may be a useful tool for exploring potential extensions of multiple testing procedures into broader settings.

Acknowledgments. We thank Wenyu Chen for helping implement the new `p-filter` algorithm, and Etienne Roquain for clarifying a point about doubly-weighted procedures. We thank Eugene Katsevich, Aditya Guntuboyina, Sujayam Saha, Larry Wasserman and Fanny Yang for relevant discus-

sions. The authors are also thankful to audience members at the Statistics departments of Stanford, Wharton, UC Davis, the St. Louis Workshop on Higher Order Asymptotics and Post-Selection Inference, and the NIPS Workshop on Adaptive Data Analysis, whose questions partly shaped this work. This work was supported in part by the Office of Naval Research under grant number W911NF-16-1-0368, the Air Force Office of Scientific Resesarch under grant number AFOSR-FA9550-14-1-0016, by NSF award DMS-1654076, and by an Alfred P. Sloan fellowship.

APPENDIX A: PROOF OF SUPER-UNIFORMITY LEMMA 1

Statement (b) of Lemma 1 follows directly from Blanchard and Roquain [9], and independently reproved by Barber and Ramdas [1]. Statement (a) with inequality (but not with equality) follows as a special case of (b), since independence is a special case of positive dependence, and the distribution of a null P_i does not change on conditioning on an independent set of p-values. Statement (c) was proved also by Blanchard and Roquain [9]. We now prove the statements (a), (d).

Statement (a). We prove the first part of Lemma 1, under the assumptions that the function $P \mapsto f(P)$ satisfies the leave-one-out property with respect to index i, and that P_i is uniformly distributed and is independent of the remaining p-values. Since $\Pr\{P_i = 0\} = 0$, we ignore this possibility in the following calculations. Since f satisfies the LOOP condition, we have

$$\frac{\mathbf{1}\{P_i \le f(P)\}}{f(P)} = \frac{\mathbf{1}\left\{P_i \le f(\widetilde{P}^{-i})\right\}}{f(\widetilde{P}^{-i})}.$$

This can be seen by separately considering what happens when the numerator on the left-hand side is zero or one.

Since P^{-i} determines $f(\widetilde{P}^{-i})$, it immediately follows that

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le f(P)\}}{f(P)} \;\middle|\; P^{-i}\right] &= \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le f(\widetilde{P}^{-i})\right\}}{f(\widetilde{P}^{-i})} \;\middle|\; P^{-i}\right] \\ &= \frac{\Pr\left\{P_i \le f(\widetilde{P}^{-i}) \;\middle|\; f(\widetilde{P}^{-i})\right\}}{f(\widetilde{P}^{-i})} \\ &= 1,\end{aligned}$$

where the last step follows since f has range $[0,1]$, and P_i is uniformly distributed and is independent of $\widetilde{P}^{-i}$— therefore, $\Pr\left\{P_i \le f(\widetilde{P}^{-i}) \;\middle|\; f(\widetilde{P}^{-i})\right\} =$

$f(\widetilde{P}^{-i})$. This concludes the proof of the super-uniformity lemma under independence and uniformity.

Statement (d). For each $\ell = 1, \ldots, m$, let ν_ℓ be a probability measure on $[0, \infty)$ chosen such that $\beta_\ell(k) = \beta_{\nu_\ell}(k) = \int_{x=0}^{k} x \, \mathsf{d}\nu_\ell(x)$, as in the definition of a reshaping function (Definition 3). Let $X_\ell \sim \nu_\ell$ be drawn independently for each $\ell = 1, \ldots, m$, and let ν be the probability measure on $[0, \infty)$ corresponding to the distribution of $Z = \prod_{\ell=1}^{m} X_\ell$. Then

$$
\begin{aligned}
c \cdot \prod_{\ell=1}^{m} \beta_\ell(f_\ell(P)) &= c \cdot \prod_{\ell=1}^{m} \left(\int_{x_\ell=0}^{f_\ell(P)} x_\ell \, \mathsf{d}\nu_\ell(x_\ell) \right) \\
&= c \cdot \int_{x_1=0}^{\infty} \cdots \int_{x_m=0}^{\infty} \left(\prod_{\ell=1}^{m} x_\ell \cdot \mathbf{1}\left\{x_\ell \le f_\ell(P)\right\} \right) \mathsf{d}\nu_m(x_m) \ldots \mathsf{d}\nu_1(x_1) \\
&= c \cdot \mathbb{E}\left[\prod_{\ell=1}^{m} \left(X_\ell \cdot \mathbf{1}\left\{X_\ell \le f_\ell(P)\right\} \right) \right] \\
&= c \cdot \mathbb{E}\left[Z \cdot \mathbf{1}\left\{X_1 \le f_1(P), \ldots, X_m \le f_m(P)\right\} \right] \\
&\le c \cdot \mathbb{E}\left[Z \cdot \mathbf{1}\left\{ Z \le \prod_{\ell=1}^{m} f_\ell(P) \right\} \right] \\
&= c \cdot \int_{z=0}^{\prod_{\ell=1}^{m} f_\ell(P)} z \, \mathsf{d}\nu(z) = c \cdot \beta_\nu \left(\prod_{\ell=1}^{m} f_\ell(P) \right).
\end{aligned}
$$

Therefore,

$$
\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le c \cdot \prod_{\ell=1}^{m} \beta_\ell(f_\ell(P))\right\}}{c \cdot \prod_{\ell=1}^{m} f_\ell(P)} \right] \le \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \le c \cdot \beta_\nu\left(\prod_{\ell=1}^{m} f_\ell(P)\right)\right\}}{c \cdot \prod_{\ell=1}^{m} f_\ell(P)} \right] \le 1,
$$

where the last step holds by Lemma 1(c).

APPENDIX B: PROOF OF INVERSE-BINOMIAL LEMMA 2

The lower bound follows immediately from Jensen's inequality, since $\mathbb{E}[Z] = 1 + b\sum_{i=1}^{n} a_i$. We split the argument for the upper bound into three cases.

Case 1: integer weights. First, suppose that all the weights a_i are integers, that is, $a_i \in \{0, 1\}$ for all i. In this case, we have $Z \sim 1 + \text{Binomial}(k, b)$, where k is the number of weights a_i that are equal to 1. A simple calculation

shows that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+\text{Binomial}(k,b)}\right] &= \sum_{z=0}^{k} \frac{1}{1+z}\binom{k}{z} b^z (1-b)^{k-z} \\
&= \frac{1}{b(1+k)} \sum_{z=0}^{k} \binom{k+1}{z+1} b^{z+1} (1-b)^{(k+1)-(z+1)} \\
&= \frac{1}{b(1+k)} \cdot \Pr\{\text{Binomial}(k+1,b) \le k\} \\
&\le \frac{1}{b(1+k)} = \frac{1}{b(1+\sum_i a_i)}.
\end{aligned}$$

Case 2: one non-integer weight. Suppose that exactly one of the weights a_i is a non-integer. Without loss of generality we can take $a_1 = \cdots = a_k = 1$, $a_{k+1} = c$, $a_{k+2} = \cdots = a_n = 0$, for some $k \in \{0, \ldots, n-1\}$ and some $c \in (0,1)$. Let $A = Z_1 + \cdots + Z_{k+1} \sim \text{Binomial}(k+1, b)$, and $Y = Z_{k+1} \sim \text{Bernoulli}(b)$. Note that $\Pr\{Y = 1 \mid A\} = \frac{A}{1+k}$. Then

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Z}\right] &= \mathbb{E}\left[\frac{1}{1+A-(1-c)Y}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{1+A-(1-c)Y} \;\middle|\; A\right]\right] \\
&= \mathbb{E}\left[\frac{1}{1+A} \cdot \Pr\{Y=0 \mid A\} + \frac{1}{c+A} \cdot \Pr\{Y=1 \mid A\}\right] \\
&= \mathbb{E}\left[\frac{1}{1+A} + \left(\frac{1}{c+A} - \frac{1}{1+A}\right) \cdot \Pr\{Y=1 \mid A\}\right] \\
&= \mathbb{E}\left[\frac{1}{1+A} + \frac{1-c}{(c+A)(1+A)} \cdot \frac{A}{1+k}\right] \\
&\le \mathbb{E}\left[\frac{1}{1+A} + \frac{1-c}{(c+1+k)(1+A)} \cdot \frac{1+k}{1+k}\right],
\end{aligned}$$

where the inequality holds since $\frac{A}{c+A} \le \frac{1+k}{1+k+c}$ because $0 \le A \le k+1$. Simplifying, we get

$$\mathbb{E}\left[\frac{1}{Z}\right] \le \mathbb{E}\left[\frac{1}{1+A}\right] \cdot \frac{2+k}{1+k+c} \le \frac{1}{b(2+k)} \cdot \frac{2+k}{1+k+c} = \frac{1}{b(1+k+c)} = \frac{1}{b(1+\sum_i a_i)},$$

where the inequality uses the fact that $\mathbb{E}\left[\frac{1}{1+\text{Binomial}(k+1,b)}\right] \le \frac{1}{b(2+k)}$ as calculated in Case 1.

Case 3: general case. Now suppose that there are at least two non-integer weights, $0 < a_i \le a_j < 1$. Let $C = \sum_{\ell \neq i,j} a_\ell Z_\ell$, then $Z = 1 + C + a_i Z_i + a_j Z_j$. Let $\alpha = \min\{a_i, 1 - a_j\} > 0$. Then

$$\mathbb{E}\left[\frac{1}{Z} \;\middle|\; C\right] = b^2 \cdot \frac{1}{1 + C + a_i + a_j} + b(1-b) \cdot \frac{1}{1 + C + a_i} + b(1-b) \cdot \frac{1}{1 + C + a_j} + (1-b)^2 \cdot \frac{1}{C}$$
$$\le b^2 \cdot \frac{1}{1 + C + a_i + a_j} + b(1-b) \cdot \frac{1}{1 + C + (a_i - \alpha)} + b(1-b) \cdot \frac{1}{1 + C + (a_j + \alpha)} + (1-b)^2 \cdot \frac{1}{C},$$

where the inequality follows from a simple calculation using the assumption that $\alpha \le a_i \le a_j \le 1 - \alpha$. Now, define a new vector of weights $\tilde{a}$ where $\tilde{a}_i = a_i - \alpha, \tilde{a}_j = a_j + \alpha$ and $\tilde{a}_\ell = a_\ell$ if $\ell \notin \{i, j\}$. Defining $\widetilde{Z} = 1 + \sum_\ell \tilde{a}_\ell Z_\ell$, the above calculation proves that $\mathbb{E}\left[\frac{1}{Z}\right] \le \mathbb{E}\left[\frac{1}{\widetilde{Z}}\right]$ (by marginalizing over C).

Note that $\sum_i a_i = \sum_i \tilde{a}_i$, but $\tilde{a}_i$ has (at least) one fewer non-integer weight. Repeating this process inductively, we see that we can reduce to the case where there is at most one non-integer weight (i.e., Case 1 or Case 2). This proves the lemma.

APPENDIX C: PROOF OF GROUP SUPER-UNIFORMITY LEMMA 3

The proof of Lemma 3(c) is straightforward, by applying Lemma 1(c). More precisely, define an augmented vector $P' = (P_1, \dots, P_n, T(P_{A_g})) \in [0,1]^{n+1}$, and define a function $f'(P') = f(P_1, \dots, P_n) = f(P)$. Since $T(P_{A_g})$ is assumed to be superuniform (since $g \in \mathcal{H}^0_{\text{grp}}$ is a null group), this means that $P'_{n+1} = T(P_{A_g})$ is superuniform, i.e. index $n+1$ is a null p-value, in the augmented vector of p-values P'. Then applying Lemma 1(c), with P' and f' in place of P and f, and with index $i = n + 1$, yields the desired bound.

Lemma 3(a) is simply a special case of Lemma 3(b) since independence is a special case of positive dependence, and conditioning on an independent set of p-values P_{-A_g} doesn't change the distribution of P_{A_g}.

For Lemma 3(b), our proof strategy will be to reduce this statement into a form where Lemma 1(b) becomes applicable. (Note that we cannot simply take the approach of our proof of Lemma 3(c), because if we define an augmented vector of p-values $P' = \big(P_1, \dots, P_n, \text{Simes}_w(P_{A_g})\big)$, we do not know if this vector is positively dependent—specifically, whether P' is PRDS on entry $P'_{n+1} = \text{Simes}_w(P_{A_g})$.)

With this aim in mind, let $\widehat{k}_g \in \{0, \dots, n_g\}$ be the number of discoveries made by the BH_w procedure when run on the p-values within group g at level $f(P)$. Then, using the connection between the Simes test and the BH

procedure, we may write

$$\mathbf{1}\{P_g \le f(P)\} = \mathbf{1}\left\{\widehat{k}_g > 0\right\} = \frac{\widehat{k}_g}{\widehat{k}_g} = \frac{\sum_{i\in A_g} \mathbf{1}\left\{P_i \le \frac{w_i \widehat{k}_g f(P)}{n_g}\right\}}{\widehat{k}_g},$$

since for the BH_w procedure at level $f(P)$, the ith p-value P_i will be rejected if and only if $P_i \le \frac{w_i \widehat{k}_g f(P)}{n_g}$. Hence, we may conclude that

$$\frac{\mathbf{1}\{P_g \le f(P)\}}{f(P)} = \frac{\sum_{i\in A_g} \mathbf{1}\left\{P_i \le \frac{w_i \widehat{k}_g f(P)}{n_g}\right\}}{\widehat{k}_g f(P)} = \frac{1}{n_g}\sum_{i\in A_g} w_i \frac{\mathbf{1}\left\{P_i \le \widetilde{f}_g(P)\right\}}{\widetilde{f}_g(P)},$$

where we have defined $\widetilde{f}_g(P) := \frac{w_i \widehat{k}_g f(P)}{n_g}$. Taking expectations on both sides and applying Lemma 1(b) immediately proves Lemma 3(b). (Specifically, we know that $P \mapsto \widehat{k}_g$ is a non-increasing function of P, and $P \mapsto f(P)$ is also assumed to be non-increasing; therefore, $\widetilde{f}_g$ is also non-increasing in P.)

Given that Lemma 3(b) is proved, the proof of Lemma 3(d) follows exactly the same argument as above, except that in the very last equation, P_i is replaced by S_i, and Lemma 3(b) is invoked in place of Lemma 1(b).

APPENDIX D: PROOF OF "MAXIMUM-CORNER" PROPOSITION 6

For each m, by definition of $\widehat{k}_m$, there is some $k_1^{(m)}, \dots, k_{m-1}^{(m)}, k_{m+1}^{(m)}, \dots, k_M^{(m)}$ such that

$$(k_1^{(m)}, \dots, k_{m-1}^{(m)}, \widehat{k}_m, k_{m+1}^{(m)}, \dots, k_M^{(m)}) \in \widehat{\mathcal{K}}\,. \tag{D.1}$$

Thus, for each $m' \neq m$, $\widehat{k}_{m'} \ge k_{m'}^{(m)}$ by definition of $\widehat{k}_{m'}$. Then

$$\widehat{\mathcal{S}}(k_1^{(m)}, \dots, k_{m-1}^{(m)}, \widehat{k}_m, k_{m+1}^{(m)}, \dots, k_M^{(m)}) \subseteq \widehat{\mathcal{S}}(\widehat{k}_1, \dots, \widehat{k}_{m-1}, \widehat{k}_m, \widehat{k}_{m+1}, \dots, \widehat{k}_M)\,,$$

because $\widehat{\mathcal{S}}(k_1, \dots, k_M)$ is a nondecreasing function of $(k_1, \dots, k_M)$, and this immediately implies

$$\widehat{\mathcal{S}}_m(k_1^{(m)}, \dots, k_{m-1}^{(m)}, \widehat{k}_m, k_{m+1}^{(m)}, \dots, k_M^{(m)}) \subseteq \widehat{\mathcal{S}}_m(\widehat{k}_1, \dots, \widehat{k}_{m-1}, \widehat{k}_m, \widehat{k}_{m+1}, \dots, \widehat{k}_M).$$

Therefore, for each layer m,

$$\sum_{g\in\widehat{\mathcal{S}}_m(\widehat{k}_1,\dots,\widehat{k}_m)} u_g^{(m)} \ge \sum_{g\in\widehat{\mathcal{S}}_m(k_1^{(m)},\dots,k_{m-1}^{(m)},\widehat{k}_m,k_{m+1}^{(m)},\dots,k_M^{(m)})} u_g^{(m)} \ge \widehat{k}_m,$$

where the second inequality holds by observation (D.1), and by definition of $\widehat{\mathcal{K}}$ as the set of feasible vectors. Since this holds for all m, this proves that $(\widehat{k}_1, \dots, \widehat{k}_M)$ is itself a feasible vector, and hence $(\widehat{k}_1, \dots, \widehat{k}_M) \in \widehat{\mathcal{K}}$.

APPENDIX E: PROOF OF THEOREM 7.1

To be able to handle all four cases of the theorem, we define a function γ_m to be the identity if we are not using reshaping (theorem statements (a,b)), or $\gamma_m = \beta_m$ if we are using reshaping (theorem statements (c,d)). We also let $\widehat{\pi}_m = 1$ and $\lambda_m = 1$ if we are not using adaptivity (as in theorem statements (b,c)), or let $\widehat{\pi}_m$ be defined as in (7.2) for theorem statements (a,d) where adaptivity is used.

Fix any partition m. Since $\Pr\{P_i = 0\} = 0$ for any $i \in \mathcal{H}^0$ by assumption, we assume that $P_i \neq 0$ for any $i \in \mathcal{H}^0$ without further mention; this assumption then implies that if $g \in \widehat{\mathcal{S}}_m(\widehat{k}_1, \ldots, \widehat{k}_M)$ for some null group $g \in \mathcal{H}^0_m$, we must have $\widehat{k}_m > 0$. We can then calculate

$$\begin{aligned}
\mathrm{FDP}^{(m)}_u(\widehat{k}_1, \ldots, \widehat{k}_M) &= \frac{\sum_{g \in \mathcal{H}^0_m} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}_m(\widehat{k}_1, \ldots, \widehat{k}_M)\right\}}{\sum_{g \in [G_m]} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}_m(\widehat{k}_1, \ldots, \widehat{k}_M)\right\}} \\
&\leq \frac{\sum_{g \in \mathcal{H}^0_m} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}_m(\widehat{k}_1, \ldots, \widehat{k}_M)\right\}}{\widehat{k}_m}, \\
&\leq \frac{\sum_{g \in \mathcal{H}^0_m} u_g^{(m)} \mathbf{1}\left\{g \in \widehat{\mathcal{S}}^{\mathrm{init}}_m(\widehat{k}_1, \ldots, \widehat{k}_M)\right\}}{\widehat{k}_m}, \\
&= \frac{\sum_{g \in \mathcal{H}^0_m} u_g^{(m)} \mathbf{1}\left\{P^m_g \leq \min\{w_g^{(m)} \frac{\alpha_m \gamma_m(\widehat{k}_m)}{\widehat{\pi}_m G_m}, \lambda_m\}\right\}}{\widehat{k}_m},
\end{aligned}$$

where the first inequality follows by definition (7.7) of the feasible set $\widehat{\mathcal{K}}$, the second follows since $\widehat{\mathcal{S}}_m(\vec{k}) \subseteq \widehat{\mathcal{S}}^{\mathrm{init}}_m(\vec{k})$ for any $\vec{k}$ by definition, and the last step uses the definition of $\widehat{\mathcal{S}}^{\mathrm{init}}_m(\vec{k})$ in (7.3a) (without reshaping, for theorem statements (a,b)) or (7.3b) (with reshaping, for theorem statements (c,d)).

Multiplying the numerator and denominator of each term by $\frac{\alpha_m w_g^{(m)}}{G_m}$, and taking expectations on both sides, we deduce that

(E.1)
$$\mathrm{FDR}^{(m)}_u \leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}^0_m} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P^m_g \leq \min\{\frac{w_g^{(m)} \alpha_m \gamma_m(\widehat{k}_m)}{\widehat{\pi}_m G_m}, \lambda_m\}\right\}}{\frac{w_g^{(m)} \alpha_m \widehat{k}_m}{G_m}}\right].$$

With these calculations in place, we now prove the four statements of the theorem.

Theorem statement (a). Define the function $f_g^{(m)}$ that maps the vector P to $\frac{w_g^{(m)}\alpha_m \widehat{k}_m}{\widehat{\pi}_m G_m}$. Note that $f_g^{(m)}$ is a nonincreasing function of P, since $\widehat{k}_m$ is a nonincreasing function of P by definition of our procedure, while $\widehat{\pi}_m$ is a nondecreasing function of P.

Returning to (E.1), we then see

$$\begin{aligned}
\mathrm{FDR}_u^{(m)} &\le \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le \min\{\frac{w_g^{(m)}\alpha_m\widehat{k}_m}{\widehat{\pi}_m G_m}, \lambda_m\}\right\}}{\widehat{\pi}_m \frac{w_g^{(m)}\alpha_m\widehat{k}_m}{\widehat{\pi}_m G_m}} \right] \\
&= \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\mathbf{1}\left\{P_g^m \le \lambda_m\right\} \cdot \frac{\mathbf{1}\left\{P_g^m \le f_g^{(m)}(P)\right\}}{\widehat{\pi}_m f_g^{(m)}(P)} \right] \\
&\le \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le f_g^{(m)}(P)\right\}}{\widehat{\pi}_m^{-g} f_g^{(m)}(P)} \right], \qquad \text{(E.2)}
\end{aligned}$$

where

$$\widehat{\pi}_m^{-g} := \frac{|u^{(m)}w^{(m)}|_\infty + \sum_{h\neq g} u_h^{(m)} w_h^{(m)} \mathbf{1}\left\{P_h^{(m)} > \lambda_m\right\}}{n(1-\lambda_m)}. \qquad \text{(E.3)}$$

Note that, on the event $P_g^m \le \lambda_m$, we have $\widehat{\pi}_m = \widehat{\pi}_m^{-g}$, which allows the inequality (E.2) above. Returning to (E.2), we now condition on $P_{-A_g^m}$ for each group g:

$$\begin{aligned}
\mathrm{FDR}_u^{(m)} &\le \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le f_g^{(m)}(P)\right\}}{\widehat{\pi}_m^{-g} f_g^{(m)}(P)} \right] \\
&= \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le f_g^{(m)}(P)\right\}}{\widehat{\pi}_m^{-g} f_g^{(m)}(P)} \,\middle|\, P_{-A_g^m} \right] \right] \\
&\overset{(i)}{=} \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le f_g^{(m)}(P)\right\}}{f_g^{(m)}(P)} \,\middle|\, P_{-A_g^m} \right] \right] \\
&\le \frac{\alpha_m}{G_m} \sum_{g\in\mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}} \right],
\end{aligned}$$

where equality (i) holds because $\widehat{\pi}_m^{-g}$ is a function of only the p-values outside of group g, i.e. of $P_{-A_g^m}$, while the last inequality holds by Lemma 3(a).

Finally, observe that independence between the different groups of partition m implies that the indicator variables $\mathbf{1}\{P_h^m > \lambda_m\}$ are independent Bernoulli's with probabilities $\leq 1 - \lambda_m$ of success. Thus, as a consequence of Lemma 2, we can prove (analogous to property (5.5)) that

$$\mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}}\right] \leq \frac{G_m}{\sum_{h \in \mathcal{H}_m^0} u_h^{(m)} w_h^{(m)}}. \tag{E.4}$$

Plugging this back into our bounds on FDR, we finally obtain

$$\begin{aligned}
\mathrm{FDR}_u^{(m)} &\leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}}\right] \\
&\leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \frac{G_m}{\sum_{h \in \mathcal{H}_m^0} u_h^{(m)} w_h^{(m)}} \\
&\leq \alpha_m.
\end{aligned}$$

Theorem statement (b). The proof of statement (b) follows the same steps as for (a), but without the need to condition on $P_{-A_g^m}$, since we do not use adaptivity. Define the function $f_g^{(m)}(P) = \frac{w_g^{(m)} \alpha_m \widehat{k}_m}{G_m}$. Then $f_g^{(m)}$ is a nonincreasing function of P, since $\widehat{k}_m$ is a nonincreasing function of P.

Returning to (E.1), as in the proof of statement (a), we calculate

$$\mathrm{FDR}_u^{(m)} \leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \leq f_g^{(m)}(P)\right\}}{f_g^{(m)}(P)}\right].$$

By Lemma 3(b), we know that $\mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \leq f_g^{(m)}(P)\right\}}{f_g^{(m)}(P)}\right] \leq 1$, and therefore

$$\mathrm{FDR}_u^{(m)} \leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)},$$

as desired.

Theorem statement (c). We now turn to proving the method under reshaping. Define $f_g^{(m)}(P) = \widehat{k}_m$, and define constant $c_g^{(m)} = \frac{w_g^{(m)} \alpha_m}{G_m}$. Returning to (E.1), as before, we calculate

$$\mathrm{FDR}_u^{(m)} \leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \leq c_g^{(m)} \cdot \beta_m\big(f_g^{(m)}(P)\big)\right\}}{c_g^{(m)} \cdot f_g^{(m)}(P)}\right].$$

By Lemma 3(c), we know that $\mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le c_g^{(m)} \cdot \beta_m\left(f_g^{(m)}(P)\right)\right\}}{c_g^{(m)} \cdot f_g^{(m)}(P)}\right] \le 1$ since P_g^m is assumed to be superuniform for any null group $g \in \mathcal{H}_m^0$. Therefore,

$$\mathrm{FDR}_u^{(m)} \le \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)}.$$

Theorem statement (d). The proof of part (d) combines the calculations of part (a) (where adaptivity is used) with part (c) (where reshaping is used). Define $f_g^{(m)} = \widehat{k}_m$ and $c_g^{(m)} = \frac{w_g^{(m)} \alpha_m}{\widehat{\pi}_m^{-g} G_m}$, where $\widehat{\pi}_m^{-g}$ is defined as in (E.3) from part (a). Note that $c_g^{(m)}$ is no longer a constant, but nonetheless, proceeding as in part (a), we can calculate

$$\mathrm{FDR}_u^{(m)} \le \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le c_g^{(m)} \cdot \beta_m\left(f_g^{(m)}(P)\right)\right\}}{\widehat{\pi}_m^{-g} \cdot c_g^{(m)} \cdot f_g^{(m)}(P)}\right].$$

Next we condition on the p-values outside the group A_g^m:

$$\begin{aligned}\mathrm{FDR}_u^{(m)} &\le \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le c_g^{(m)} \cdot \beta_m\left(f_g^{(m)}(P)\right)\right\}}{\widehat{\pi}_m^{-g} \cdot c_g^{(m)} \cdot f_g^{(m)}(P)} \,\middle|\, P_{-A_g^m}\right]\right] \\ &= \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}} \cdot \mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le c_g^{(m)} \cdot \beta_m\left(f_g^{(m)}(P)\right)\right\}}{c_g^{(m)} \cdot f_g^{(m)}(P)} \,\middle|\, P_{-A_g^m}\right]\right],\end{aligned}$$

where the last step holds since $\widehat{\pi}_m^{-g}$ is a function of $P_{-A_g^m}$.

Finally, we will apply Lemma 3(c) to show that each of these conditional expected values is ≤ 1. Of course, the subtlety here is that we must condition on $P_{-A_g^m}$. To do so, note that, after fixing $P_{-A_g^m}$, the function $f_g^{(m)}(P)$ can be regarded as a function of only the remaining unknowns (i.e. of $P_{A_g^m}$), and is still non-increasing; the value $c_g^{(m)}$ is now a constant; and $P_g^m = T_g^m(P_{A_g^m})$ is indeed superuniform since, due to the independence of $P_{A_g^m}$ from $P_{-A_g^m}$, its distribution has not changed. Therefore, we can apply Lemma 3(c) (with the random vector $P_{A_g^m}$ in place of P, while $P_{-A_g^m}$ is treated as constant), to see that $\mathbb{E}\left[\frac{\mathbf{1}\left\{P_g^m \le c_g^{(m)} \cdot \beta_m\left(f_g^{(m)}(P)\right)\right\}}{c_g^{(m)} \cdot f_g^{(m)}(P)} \,\middle|\, P_{-A_g^m}\right] \le 1$, and therefore,

$$\mathrm{FDR}_u^{(m)} \le \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \mathbb{E}\left[\frac{1}{\widehat{\pi}_m^{-g}}\right].$$

Finally, we need to bound $\widehat{\pi}_m^{-g}$. As in the proof of part (a), we see that the indicator variables $\mathbf{1}\{P_h^m > \lambda_m\}$ are independent, since $P_h^m = T_h^m(P_{A_h^m})$, and the sets of p-values $P_{A_h^m}$ are assumed to be independent from each other. Furthermore, since $T_h^m(P_{A_h^m})$ is assumed to be a valid p-value, i.e. superuniform for any $h \in \mathcal{H}_m^0$, this means that the variable $\mathbf{1}\{P_h^m > \lambda_m\}$ is Bernoulli with chance $\geq 1 - \lambda_m$ of success. Therefore, the bound (E.4) calculated in the proof of part (a) holds here as well, and so

$$\mathrm{FDR}_u^{(m)} \leq \frac{\alpha_m}{G_m} \sum_{g \in \mathcal{H}_m^0} u_g^{(m)} w_g^{(m)} \frac{G_m}{\sum_{h \in \mathcal{H}_m^0} u_h^{(m)} w_h^{(m)}} = \alpha_m.$$

This concludes the proof of all four parts of Theorem 7.1.

APPENDIX F: PROOF OF PROPOSITION 7

First we introduce some notation: let $(k_1^{(s)}, \ldots, k_M^{(s)})$ be the vector after the sth pass through the algorithm. We prove that $k_m^{(s)} \geq \widehat{k}_m$ for all m, s, by induction. At initialization, $k_m^{(0)} = G_m \geq \widehat{k}_m$ for all m. Now suppose that $k_m^{(s-1)} \geq \widehat{k}_m$ for all m; we now show that $k_m^{(s)} \geq \widehat{k}_m$ for all m.

To do this, consider the m-th layer of the s-th pass through the algorithm. Before this stage, we have vectors $k_1^{(s)}, \ldots, k_{m-1}^{(s)}, k_m^{(s-1)}, k_{m+1}^{(s-1)}, \ldots, k_M^{(s-1)}$, and we now update $k_m^{(s)}$. Applying induction also to this inner loop, and assuming that $k_{m'}^{(s)} \geq \widehat{k}_{m'}$ for all $m' = 1, \ldots, m-1$, we can now prove that $k_m^{(s)} \geq \widehat{k}_m$. By definition of the algorithm,

$$k_m^{(s)} = \max_{k'_m \in \{0,1,\ldots,G_m\}} \left\{ k'_m : \sum_{g \in \widehat{\mathcal{S}}_m(k_1^{(s)}, \ldots, k_{m-1}^{(s)}, k'_m, k_{m+1}^{(s-1)}, \ldots, k_M^{(s-1)})} u_g^{(m)} \geq k'_m \right\}. \tag{F.1}$$

Since $k_{m'}^{(s)} \geq \widehat{k}_{m'}$ for all $m' = 1, \ldots, m-1$, and $k_{m'}^{(s-1)} \geq \widehat{k}_{m'}$ for all $m' = m+1, \ldots, M$, we have

$$\sum_{g \in \widehat{\mathcal{S}}_m(k_1^{(s)}, \ldots, k_{m-1}^{(s)}, \widehat{k}_m, k_{m+1}^{(s-1)}, \ldots, k_M^{(s-1)})} u_g^{(m)} \geq \sum_{g \in \widehat{\mathcal{S}}_m(\widehat{k}_1, \ldots, \widehat{k}_{m-1}, \widehat{k}_m, \widehat{k}_{m+1}, \ldots, \widehat{k}_M)} u_g^{(m)}$$

since $\widehat{\mathcal{S}}_m(\vec{k})$ is a nondecreasing function of $\vec{k}$ by definition. The right-hand side of this expression is in turn $\geq \widehat{k}_m$ by definition of $(\widehat{k}_1, \ldots, \widehat{k}_M)$ being a feasible vector. Therefore, $\widehat{k}_m$ is in the feasible set for Eq. (F.1), and so we must have $k_m^{(s)} \geq \widehat{k}_m$. By induction, this is then true for all s, m, as desired.

Now suppose that the algorithm stabilizes at $(k_1^{(s)}, \ldots, k_M^{(s)})$, after s full passes. After completing the mth layer of the last pass through the algorithm, we had vectors $k_1^{(s)}, \ldots, k_m^{(s)}, k_{m+1}^{(s-1)}, \ldots, k_M^{(s-1)}$; however, since the algorithm stops after the sth pass, this means that $k_{m'}^{(s-1)} = k_{m'}^{(s)}$ for all m'. Using this observation in the definition of $k_m^{(s)}$, we see that

$$\sum_{g \in \widehat{\mathcal{S}}_m(k_1^{(s)}, \ldots, k_{m-1}^{(s)}, k_m^{(s)}, k_{m+1}^{(s)}, \ldots, k_M^{(s)})} u_g^{(m)} \geq k_m^{(s)}.$$

This means that $(k_1^{(s)}, \ldots, k_M^{(s)}) \in \widehat{\mathcal{K}}$, and so $k_m^{(s)} \leq \widehat{k}_m$ for all m by the definition of $\widehat{k}_1, \ldots, \widehat{k}_m$ and Proposition 6. But by the induction above, we also know that $k_m^{(s)} \geq \widehat{k}_m$ for all m at any iteration s; and this completes the proof.

APPENDIX G: PROPERTIES OF DOTFRACTIONS

In this section we verify that "dotfractions" satisfy many of the same properties as ordinary fractions, and thus the notation $\frac{a}{\cdot\cdot b}$ can be safely used throughout the proofs of our main results. In all the following, the property will be shown to hold assuming that all dotfractions appearing in its equation or inequality are well defined. Hence, throughout, we assume that the various properties are only used if all of the dotfractions in the expression are defined—that is, we may use these properties only if we never have $\frac{a}{\cdot\cdot b}$ with $a \neq 0$ and $b = 0$. As a side note, observe that in the paper, we always use $\frac{a}{\cdot\cdot b}$ when $a, b \geq 0$ only.

1. Comparing two fractions:

   (G.1) If $a \geq b \geq 0$ and $c \geq 0$, then $\frac{a}{\cdots c} \geq \frac{b}{\cdot\cdot c}$, and $\frac{c}{\cdots a} \leq \frac{c}{\cdot\cdot b}$.

   To prove the first bound, if $c > 0$ then this reduces to $\frac{a}{c} \geq \frac{b}{c}$, while if $c = 0$ then we must have $a = b = 0$ (since, otherwise, $\frac{a}{\cdot\cdot c}$ and $\frac{b}{\cdot\cdot c}$ would be undefined) and so $\frac{a}{\cdot\cdot c} = \frac{b}{\cdot\cdot c} = 0$. To prove the second bound, if $b > 0$, then this reduces to $\frac{c}{a} \leq \frac{c}{b}$, while if $b = 0$ then we must have $c = 0$ (since, otherwise, $\frac{c}{\cdot\cdot b}$ would be undefined), in which case $\frac{c}{\cdot\cdot a} = \frac{c}{\cdot\cdot b} = 0$.
2. Comparing against a scalar:

   (G.2) If $c \geq 0$ and $a \geq \frac{b}{\cdot\cdot c}$ then $ac \geq b$.

   To prove this, if $c \neq 0$ then we have $a \geq \frac{b}{c}$, while if $c = 0$ then we must have $b = 0$ (so that $\frac{b}{\cdot\cdot c}$ is not undefined), and so $ac \geq b$ is trivially true as both sides equal zero.

3. Adding numerators:

$$\text{(G.3)} \qquad \text{For any } a,b,c, \quad \frac{a}{c} + \frac{b}{c} = \frac{a+b}{c}.$$

To prove this, if $c \neq 0$ then this reduces to $\frac{a}{c} + \frac{b}{c} = \frac{a+b}{c}$, while if $c = 0$ then we must have $a = b = 0$ (otherwise the dotfractions are undefined), and so the left- and right-hand sides both equal zero.

4. Multiplying fractions:

$$\text{(G.4)} \qquad \text{For any } a,b,c,d, \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

To prove this, if $b, d \neq 0$ then this reduces to $\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$, while if $b = 0$ or $d = 0$, then either $a = 0$ or $c = 0$ (otherwise $\frac{a}{b}$ or $\frac{c}{d}$ would be undefined), and so the left- and right-hand sides both equal zero.

5. Cancelling nonzero factors :

$$\text{(G.5)} \qquad \text{If } c \neq 0 \text{ then for any } a,b, \quad \frac{ac}{bc} = \frac{a}{b}.$$

To see why, we simply apply (G.4) with $d = c$ (noting that, with the assumption $c \neq 0$, we have $\frac{c}{c} = 1$).

6. Multiplying by a scalar:

$$\text{(G.6)} \qquad \text{For any } a,b,c, \quad c \cdot \frac{a}{b} = \frac{ac}{b}.$$

To see why, if $b \neq 0$ then this reduces to $c \cdot \frac{a}{b} = \frac{ac}{b}$, while if $b = 0$ then we must have $a = 0$ so that $\frac{a}{b}$ is not undefined, and so the left- and right-hand sides are both zero.

While the above properties all carry over from fractions to dotfractions, there are some settings where familiar manipulations with fractions may no longer be correct. For example, $\frac{a}{b} \neq \frac{ac}{bc}$ when $a, b \neq 0$ while $c = 0$. Relatedly, we cannot add fractions in the usual way, i.e. $\frac{a}{b} + \frac{c}{d}$ may not be equal to $\frac{ad+bc}{bd}$; this fails because implicitly we would be assuming that $\frac{a}{b} = \frac{ad}{bd}$ and $\frac{c}{d} = \frac{bc}{bd}$ in order to make the two denominators the same, which may fail if $d = 0$ or if $b = 0$.

REFERENCES

[1] Rina Foygel Barber and Aaditya Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[2] Yoav Benjamini and Marina Bogomolov. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):297–318, 2014.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[4] Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[5] Yoav Benjamini and Wei Liu. A distribution-free multiple test procedure that controls the false discovery rate. *Tel Aviv, Tel Aviv University*, 1999.

[6] Yoav Benjamini and Wei Liu. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82(1):163–170, 1999.

[7] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of statistics*, 29(4):1165–1188, 2001.

[8] Yoav Benjamini, Abba Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93:491–507, 2006.

[9] Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

[10] Gilles Blanchard and Étienne Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(Dec):2837–2871, 2009.

[11] Gunnar Eklund. Massignifikansproblemet. In *Unpublished seminar papers, Uppsala University Institute of Statistics*, volume 1963, 1961.

[12] Gunnar Eklund and Paul Seeger. Massignifikansanalys. *Statistisk tidskrift*, 5:355–365, 1965.

[13] Livio Finos and Luigi Salmaso. FDR-and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference*, 137(12):3859–3870, 2007.

[14] Ruben Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250, 1969.

[15] Yulia Gavrilov, Yoav Benjamini, and Sanat K Sarkar. An adaptive step-down procedure with proven FDR control under independence. *The Annals of Statistics*, pages 619–629, 2009.

[16] Christopher Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[17] Nicholas Heard and Patrick Rubin-Delanchy. Choosing between methods for combining p-values. *arXiv preprint arXiv:1707.06897*, 2017.

[18] Philipp Heesen and Arnold Janssen. Dynamic adaptive multiple tests with finite sample FDR control. *Journal of Statistical Planning and Inference*, 168:38–51, 2016.

[19] Yosef Hochberg and Uri Liberman. An extended Simes' test. *Statistics & Probability Letters*, 21(2):101–105, 1994.

[20] G Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrische Zeitschrift*, 25(5):423–430, 1983.

[21] James Hu, Hongyu Zhao, and Harrison Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491), 2010.

[22] Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

[23] Eugene Katsevich and Chiara Sabatti. Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *arXiv preprint arXiv:1706.09375*, 2017.

[24] Erich Leo Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, pages 1137–1153, 1966.

[25] D Morgenstern. Berechnung des maximalen signifikanzniveaus des testes lehneh0 ab,

wennk untern gegebenen tests zur ablehnung führen. *Metrika*, 27(1):285–286, 1980.
[26] Joseph Romano and Azeem Shaikh. On stepdown control of the false discovery proportion. In *Optimality*, pages 33–50. Institute of Mathematical Statistics, 2006.
[27] Joseph Romano, Azeem Shaikh, and Michael Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1):1–25, 2011.
[28] B Rüger. Das maximale signifikanzniveau des tests:lehneh o ab, wennk untern gegebenen tests zur ablehnung führen. *Metrika*, 25(1):171–178, 1978.
[29] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, pages 623–632, 1982.
[30] Sanat Sarkar. On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 135–168, 2008.
[31] Sanat Sarkar. Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference*, 138(4):1072–1084, 2008.
[32] Paul Seeger. A note on a method for the analysis of significances en masse. *Technometrics*, 10(3):586–593, 1968.
[33] John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
[34] Eckart Sonnemann. *Allgemeine Lösungen multipler Testprobleme*. Universität Bern. Institut für Mathematische Statistik und Versicherungslehre, 1982.
[35] Eckart Sonnemann. Allgemeine lösungen multipler testprobleme. edv in medizin und biologie 1982, 13 (4): 120–128. english version of the original article with minor corrections by finner h: General solutions to multiple testing problems. *Biom J*, 50 (5):641–656, 2008.
[36] Eckart Sonnemann and Helmut Finner. Vollständigkeitssätze für multiple testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 121–135. Springer, 1988.
[37] John Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.
[38] John Storey, Jonathan Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.
[39] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.
[40] John Wilder Tukey. *The Problem of Multiple Comparisons: Introduction and Parts A, B, and C*. Princeton University, 1953.
[41] Vladimir Vovk. Combining p-values via averaging. *arXiv preprint arXiv:1212.4966*, 2012.
[42] Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008.

Aaditya Ramdas
Departments of Statistics and EECS
University of California, Berkeley
E-mail: aramdas@berkeley.edu

Rina Foygel Barber
Department of Statistics
University of Chicago
E-mail: rina@uchicago.edu

Martin J. Wainwright
Departments of Statistics and EECS
University of California, Berkeley
E-mail: wainwrig@berkeley.edu

Michael I. Jordan
Departments of Statistics and EECS
University of California, Berkeley
E-mail: jordan@berkeley.edu