False discovery rate control with e-values

Ruodu Wang* Aaditya Ramdas†

December 14, 2021

Abstract

E-values have gained attention as potential alternatives to p-values as measures of uncertainty, significance and evidence. In brief, e-values are realized by random variables with expectation at most one under the null; examples include betting scores, (point null) Bayes factors, likelihood ratios and stopped supermartingales. We design a natural analog of the Benjamini-Hochberg (BH) procedure for false discovery rate (FDR) control that utilizes e-values, called the e-BH procedure, and compare it with the standard procedure for p-values. One of our central results is that, unlike the usual BH procedure, the e-BH procedure controls the FDR at the desired level—with no correction—for any dependence structure between the e-values. We illustrate that the new procedure is convenient in various settings of complicated dependence, structured and post-selection hypotheses, and multi-armed bandit problems. Moreover, the BH procedure is a special case of the e-BH procedure through calibration between p-values and e-values. Overall, the e-BH procedure is a novel, powerful and general tool for multiple testing under dependence, that is complementary to the BH procedure, each being an appropriate choice in different applications.

Keywords: multiple testing, FDR, p-values, betting scores, supermartingales

1 Introduction

We study procedures for controlling the false discovery rate (FDR) as in Benjamini and Hochberg (1995). We will encounter p-values in this paper, but focus more on e-values and e-tests as in Vovk and Wang (2021). We use "e-value" as an abstract umbrella term which encompasses betting scores, likelihood ratios, and stopped supermartingales, which appear in the recent literature, e.g., Shafer (2021); Grünwald et al. (2020); Wasserman et al. (2020) and Howard et al. (2020, 2021). (Since Bayes factors (Kass and Raftery, 1995) for point nulls are also e-values, one may hope that e-values appeal to adherents of frequentist, Bayesian and game-theoretic foundations of probability, but this paper abstains from further philosophical discussion, treating e-values as a useful technical tool that arises naturally in many situations.)

Throughout, let $H_1, \dots, H_K$ be K hypotheses, and write $\mathcal{K} = \{1, \dots, K\}$. Let the true (unknown) data-generating probability measure be denoted by $\mathbb{P}$. For each $k \in \mathcal{K}$, it is useful to think of hypothesis H_k as implicitly defining a set of joint probability measures, and H_k is called a true null hypothesis if $\mathbb{P} \in H_k$.

Following Vovk and Wang (2021), a *p-variable* P is a random variable that satisfies $\mathbb{P}(P \le \alpha) \le \alpha$ (often with equality) for all $\alpha \in (0, 1)$. In other words, a p-variable is stochastically larger than $\mathrm{U}[0, 1]$ (values of P larger than 1 can be treated as 1). An *e-variable* E is a $[0, \infty]$-valued random variable satisfying $\mathbb{E}[E] \le 1$. E-variables are often obtained from stopping an *e-process* $(E_t)_{t \ge 0}$, which is a nonnegative stochastic process adapted to a pre-specified filtration such that $\mathbb{E}[E_\tau] \le 1$ for any stopping time τ (an example would be a supermartingale with initial value 1).

*Department of Statistics and Actuarial Science, University of Waterloo. E-mail: wang@uwaterloo.ca.
†Departments of Statistics and Machine Learning, Carnegie Mellon University. E-mail: aramdas@cmu.edu.

Let $\mathcal{N} \subseteq \mathcal{K}$ be the set of indices of true null hypotheses, which is unknown to the decision maker, and K_0 be the number of true null hypotheses, thus the cardinality of $\mathcal{N}$. The ratio K_0/K may be close to 1, meaning that the signals are sparse.

For reasons to use e-values over p-values, see Shafer (2021), Vovk and Wang (2021) and Grünwald et al. (2020); however, we summarize our own perspectives later in Section 2. We consider two settings of testing multiple hypotheses, and sometimes convert between them:

1. For each $k \in \mathcal{K}$, H_k is associated with p-value p_k, which is a realization of a random variable P_k. If $k \in \mathcal{N}$, then P_k is a p-variable.

2. For each $k \in \mathcal{K}$, H_k is associated with e-value e_k, which is a realization of a random variable E_k. If $k \in \mathcal{N}$, then E_k is an e-variable.

A *p-testing procedure* $\mathcal{D} : [0,1]^K \to 2^{\mathcal{K}}$ (resp. an *e-testing procedure* $\mathcal{D} : [0,\infty]^K \to 2^{\mathcal{K}}$) gives the indices of rejected hypotheses based on observed p-values (resp. e-values). We tacitly require that all testing procedures are Borel functions. The terms "p-values/e-values" refer to both the random variables and their realized values; these should be clear from the context.

Let $\mathcal{D}$ be a p-testing procedure or an e-testing procedure. The rejected hypotheses by $\mathcal{D}$ are called discoveries. We write $F_{\mathcal{D}} := |\mathcal{D} \cap \mathcal{N}|$ as the number of true null hypotheses that are rejected (i.e., false discoveries), and $R_{\mathcal{D}} := |\mathcal{D}|$ as the total number of discoveries. The value of interest is $F_{\mathcal{D}}/R_{\mathcal{D}}$, called the false discovery proportion (FDP), which is the ratio of the number of false discoveries to that of all claimed discoveries, with the convention $0/0 = 0$ (i.e., FDP is 0 if there is no discovery). Since both quantities $F_{\mathcal{D}}$ and $R_{\mathcal{D}}$ are random, Benjamini and Hochberg (1995) proposed to control FDR, which is the expected value of FDP, that is, $\mathrm{FDR}_{\mathcal{D}} := \mathbb{E}[F_{\mathcal{D}}/R_{\mathcal{D}}]$, where the expected value is taken under the true probability. Other ways of controlling the false discovery rather than the FDR are studied by, for instance, Genovese and Wasserman (2004, 2006) and Goeman and Solari (2011) with p-values, and Vovk and Wang (2019) with e-values. We focus on controlling FDR in this paper for its popularity in modern sciences.

For those unfamiliar with e-values, we offer a couple of quick remarks. First, an e-variable E can be converted to a p-variable $P = 1/E$; the validity of P can be checked using Markov's inequality. However, since Markov's inequality is not tight, such a translation is not airtight. A p-variable P can also be converted to an e-variable, but this is more complex; one simple example is to set $E = P^{-1/2} - 1$ (and more generally $f(P)$ is an e-variable if $\int_0^1 f(u)\,\mathrm{d}u \le 1$). Importantly, small p-values correspond to large e-values, so we will reject e-values above some threshold.

A brief summary of our contributions

For $k \in \mathcal{K}$, let $e_{[k]}$ be the k-th order statistic of $e_1, \dots, e_K$, from *the largest to the smallest*. We design a simple e-value analog of the standard BH procedure, which will be called the *base e-BH procedure*. For $\alpha > 0$, define the e-testing procedure $\mathcal{G}(\alpha) : [0,\infty]^K \to 2^{\mathcal{K}}$ which rejects hypotheses with the largest k_e^* e-values, where

$$k_e^* = \max\left\{k \in \mathcal{K} : \frac{k e_{[k]}}{K} \ge \frac{1}{\alpha}\right\}, \tag{1}$$

with the convention $\max(\varnothing) = 0$, and accepts the rest.

The astute reader may note that since $1/e_k$ is a valid p-value, the base e-BH procedure is simply the BH procedure applied to the corresponding p-values. Hence, at first glance, apparently nothing is gained with the e-value viewpoint. However, one of our central results is the surprising property that

> the base e-BH procedure controls FDR at level α even under unknown *arbitrary dependence* between the e-values.

It is well known that such a statement is not true for general arbitrarily dependent p-values. Moreover, the *full e-BH procedure*, to be specified in Section 4, involves a pre-screening of e-values, allowing us to "boost" them up by a factor before feeding them to the base e-BH procedure.

We also derive several important, but secondary, results. These include an analysis of self-consistent e-testing procedures, the fact that known results on FDR control for BH and BY (Benjamini and Yekutieli, 2001) procedures can be derived as special cases of the e-BH procedure, and the optimality of e-BH procedure amongst all FDR controlling procedures acting on arbitrarily dependent e-values.

Finally, the paper contains several other tertiary contributions, including a mathematical study of how conservative the e-BH procedure is under arbitrary dependence, independence and PRDS (positive regression dependence on a subset). We provide a few interesting motivating examples along the way: identifying skillful traders (finance) and identifying promising slot machines by adaptive sampling (multi-armed bandits). Simulation results on the e-BH procedure are produced in Section 7 and Appendix A. A real-data analysis on returns of cryptocurrencies is produced in Section 8. Proofs of most results are put in Appendix B.

Before delving into the above, it seems critical to first answer a motivating question: why use e-values in the first place?

2 When might one prefer e-values over p-values?

Given that the Benjamini-Hochberg procedure using p-values is so widely established and accepted in the sciences, one must first answer the question of why one should care about FDR control (or multiple testing) with e-values. This is a philosophical, subjective, question and the opinions of the authors (and readers) are complementary to the contributions of this paper. For example, in his upcoming discussion paper, Shafer (2021) eloquently provides various arguments for the use of "betting scores", which are e-values with a game-theoretic interpretation, as measures of evidence and uncertainty in the scientific enterprise while Grünwald et al. (2020) find them to be a middle ground between Bayesian and frequentist (and financial, when combined with the earlier view) interpretations of evidence. Other authors have also already weighed in on the subject, but rather than deferring to them, we wish to offer our own perspectives in light of recent research.

Below, we list some situations when one may potentially prefer to use e-values over p-values for non-philosophical, purely statistical, reasons. To avoid confusion, we note that *the following discussion relates to the construction of a single e-value or a single p-value*—in other words, it applies equally to single and multiple hypothesis testing.

1. **High-dimensional asymptotics**. One of the most classical ways to compute p-values is to use the asymptotic distribution of the likelihood ratio test statistic, as given by Wilks' theorem (Wilks (1938)). However, the correctness of Wilks' theorem is typically justified when the dimensionality d of the data remains fixed, and the sample size n tends to infinity. There are several results on the high-dimensional asymptotics of likelihood ratios, but the resulting p-value relies on the practitioner making assumptions on the relative scalings of n and d; see Jiang and Yang (2013) for example. In contrast, the likelihood ratio for a point null hypothesis is a valid e-value in finite samples, meaning that its expectation equals one under the null regardless of d or n. The same holds for mixtures (over the alternative) of likelihood ratios as well.

2. **Irregular (composite) models**. There are many composite null hypothesis testing problems for which we may know of no direct way to construct a valid p-value even under low-dimensional asymptotics—this could happen because the model is singular or irregular and Wilks' theorem fails to hold (Drton, 2009) and in such cases the validity of the bootstrap is also typically unknown. Recently, the *split* likelihood ratio statistic (Wasserman et al., 2020) was developed, along with several other variants, and shown to yield an e-value under no assumptions on the composite null, or on d and n. Examples of new settings in which one can now construct e-values under no regularity assumptions include mixtures (e.g., testing if data comes from a mixture of $\leq k$ versus $> k$ components), shape-constraints (e.g., log-concavity), dependence structures (e.g., multivariate total positivity), and several latent variable models.

3. **Sequential inference**. The correctness of Wald's sequential likelihood ratio test is based on noting that the likelihood ratio process (L_t) is a nonnegative martingale with initial value one. Thus, at any stopping time τ, the optional stopping theorem implies that L_τ is an e-value since its expectation is at most one; in other words, (L_t) is an e-process. Moving beyond parametric settings, sequential inference is often enabled by designing nonparametric supermartingales (Howard et al., 2020) that immediately yield e-values at the stopping time. These have been called test martingales in prior work (Shafer et al., 2011). In summary, e-values arise very naturally in sequential inference as stopped e-processes. For point nulls, e-processes can always be dominated by nonnegative martingales (meaning they can be increased to a martingale without threat to validity). But for composite nulls, there is a big and important difference between e-processes, supermartingales and martingales. See Ramdas et al. (2020, 2021) for details.

4. **Accumulation of information and evidence**. Suppose the ultimate goal of a scientist is to either reject (with level α) or not reject a given hypothesis *in a single run*. Then, the most powerful method is to reject a p-value p no larger than α. This testing procedure can also be achieved by a simple e-value $e := \mathbb{1}_{\{p \leq \alpha\}}/\alpha$, using a threshold $1/\alpha$ (called an all-or-nothing bet by Shafer (2021)); in fact, no other e-values would be more powerful in general. The situation becomes quite different when the hypothesis (if seen as promising by the first experiment) will be tested with future evidence and possibly by other scientists: the all-or-nothing e-value carries little information for the next studies, whereas a generic e-value can provide a continuum of evidence strength. Even if the first e-value is moderate (e.g., $e_1 = 3$), it is useful and can be easily merged into the next experiment (Vovk and Wang, 2021), for example by simply multiplying the two e-values (this is valid if $\mathbb{E}[E_2|E_1 = e_1] \leq 1$ under the null) or averaging them (always valid). On the other hand, a p-value of 0.1 is difficult to use for future studies that may depend on data from the current one, and may be discarded (or worse, "p-hacked") due to the all-or-nothing nature of the p-test or the difficulty of merging p-values under such sequential dependence as above. Hence, for a *dynamic flow* of experiments, common in modern sciences, it is beneficial to track and report e-values — see Shafer (2021) and Grünwald et al. (2020), as well as their counterpart in estimation, confidence sequences (Howard et al., 2021).

5. **Robustness to misspecfication**. In genetics, it is not uncommon to encounter p-values with astronomically small values (like 10^{-20}), or sometimes point masses near the value one, even though the sample sizes may not intuitively support such extreme evidence. This is often (but not always) reflective of utilizing a model that is not perfectly specified. The validity of p-values is quite sensitive to model misspecification, because they utilize the entire (hypothesized) distribution of the test statistic. In contrast, e-values can be constructed without over-reliance on fine-grained tail information, and so they are typically more robust than p-values to misspecification (but less powerful under perfect specification). For example, instead of assuming that the data X is Gaussian to build a p-value, we may instead assume that it is symmetric about the origin (under no further moment constraints), in which case $\exp(\lambda X - \lambda^2 X^2/2)$ is a valid e-value for any $\lambda \in \mathbb{R}$. The proof, by Bahadur and Eaton, was communicated by Efron (1969). (However, it is inadmissible; admissible improvements can be found in Ramdas et al. (2020).)

6. **Robustness to dependence**. To construct a single valid p-value, it is often assumed (for convenience of deriving limiting distributions) that the underlying observations are independent, and indeed their validity is often hurt if this assumption is violated. However, we can often construct an e-value quite easily in settings where the observations are dependent, and this is because we are requiring less of an e-value (just bounded in expectation) than of a p-value (knowledge of its whole distribution). As one example, suppose we observe non-negative data $X_1, \ldots, X_n$, and we wish to test $H_0 : \mathbb{E}[X_i] \leq \mu$ for all i. Then, $(X_1 + \cdots + X_n)/(n\mu)$ is an e-value for any dependence structure of $X_1, \ldots, X_n$, and does not require making any distributional assumption. In addition, combination of multiple p-values relies heavily on dependence assumptions (Vovk and Wang, 2020), unlike that of e-values (Vovk and Wang, 2021). In the

field of risk management, analysis of risks with unknown or complicated dependence has recently been an active topic; see e.g., (Embrechts et al., 2015).

Thus, there exist many settings when one can, and should, use e-values to quantify evidence against a null. (However, there of course remain innumerable situations in which p-values are perfectly reasonable choices.)

To reflect some of the above points, in a concrete example below, we will see that a complicated dependence structure appears both within experiment and across experiments, and one can easily construct useful e-values in such situations, with almost no model assumptions on the test statistics.

Example 1. Suppose there are K traders (or machines), and a researcher is interested in knowing which ones are skillful (or useful). This is a classic problem in finance; see e.g., Barras et al. (2010) in the context of detecting mutual fund skills. For $k = 1, \dots, K$, the null hypothesis H_k is that trader k is not skillful, meaning that they make no profit on average (without loss of generality we can assume the market risk-free return rate is 0). The nonnegative random variables $X_{k,1}, \dots, X_{k,n}$ are the monthly realized performance (i.e., the ratio of payoff to investment; $X_{k,j} > 1$ presents a profit and $X_{k,j} < 1$ means a loss) of agent k from month 1 to month n. The no-skill null hypothesis is $\mathbb{E}[X_{k,j} \mid \mathcal{F}_{j-1}] \le 1$ for $j = 1, \dots, n$, where the σ-field $\mathcal{F}_t$ represents the available market information up to time $t \in \{0, \dots, n\}$, and we naturally assume that $(X_{k,t})_t$ is adapted to $(\mathcal{F}_t)_t$.

Since the agents are changing investment strategies over time and all strategies depend on the financial market evolution, there is complicated serial dependence within $(X_{k,1}, \dots, X_{k,n})$ for single k, as well as cross dependence among agents $k = 1, \dots, K$. Because of the complicated serial dependence and the lack of distributional assumptions of the performance data, it is difficult to obtain useful p-values for these agents. Nevertheless, we can easily obtain useful e-values: for instance, $E_k = \prod_{j=1}^n X_{k,j}$ is a valid e-value, as well as any mixture of U-statistics of $X_{k,1}, \dots, X_{k,n}$, including the mean and the product ($X_{k,1}, \dots, X_{k,n}$ are called sequential e-values by (Vovk and Wang, 2021), who also propose methods to merge them into one e-value). Moreover, the obtained e-values $E_1, \dots, E_K$ are dependent in a complicated way. Even if these e-values are not very large, they can be useful for other studies on these traders. We return to a similar example in the simulations (Section 7).

As a side note, other sophisticated e-values can also be constructed for this problem using exponential self-normalized processes (Howard et al., 2020). For example, define $\psi(\lambda) := -\lambda - \log(1-\lambda)$. Then, $E_k = \prod_{j=1}^n \exp\left(\lambda(X_{k,j} - 1) - \psi(\lambda)(X_{k,j} - 1)^2\right)$ for for any $\lambda \in [0, 1)$. Note that $\lim_{\lambda \to 0} \frac{\psi(\lambda)}{\lambda^2/2} = 1$, so this e-value captures subGaussian-like right tails of $\sum_{j=1}^n X_{k,j} - 1$ relative to its empirical variance $\sum_{j=1}^n (X_{k,j} - 1)^2$. To avoid picking a fixed λ, one can instead simply use a mixture over λ, for example using a Gamma distribution peaking near 0 (Howard et al., 2021). One can also pick λ_j predictably at each step. These observations form the basis of all time-uniform Empirical-Bernstein inequalities (Howard et al., 2021; Waudby-Smith and Ramdas, 2020a,b).

To summarize, there are several reasons to work with e-values: they arise naturally in sequential settings, we know how to construct e-values in settings where we do not know how to construct p-values, and e-values can be more robust to misspecification or uncertain asymptotics in high-dimensional settings. Of course, there are also many reasons *not* to work with e-values: in particular, p-values will often yield more "powerful" (single/multiple) tests when the underlying modeling assumptions are true. Thus, e-values and p-values can also be viewed as picking different points on the validity-power curve: if one takes more risks with validity (p-values), one can gain more power, but if one questions one or more modeling assumptions, e-values may provide a safer choice.

Remark 1. Recall that we can always convert an e-value e into a p-value $p = 1/e$. One can also transform a p-value p into an e-value $e = f(p)$ using "calibrators"; recalling the definition by Shafer et al. (2011), a calibrator is a decreasing function $f : [0, 1] \to [0, \infty)$ such that

$$\int_0^1 f(u)\, \mathrm{d}u = 1. \tag{2}$$

If the only known way to construct an e-value is by calibrating a bonafide p-value, then e-BH will typically not be more powerful than BH. Nevertheless, in some of our examples, we will assume that we can construct p-values, and calibrate them to e-values and study the resulting e-BH procedure, in order to understand the difference between the two approaches. In contrast, if the only known way to construct a p-value is by inverting an e-value, then working with e-values directly is beneficial because the e-BH procedure is robust to arbitrary dependence without corrections, unlike the BH procedure.

3 Recap: the BH procedure

We briefly review the BH and BY procedures (Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001)) for FDR control, as the benchmark for our discussions.

To discuss the dependence structure among p-values and e-values, we rely on the notion of *positive regression dependence on a subset (PRDS)* of Benjamini and Yekutieli (2001), flipped when imposed on e-values. A set $A \subseteq \mathbb{R}^K$ is said to be *decreasing* (resp. *increasing*) if $\mathbf{x} \in A$ implies $\mathbf{y} \in A$ for all $\mathbf{y} \le \mathbf{x}$ (resp. all $\mathbf{y} \ge \mathbf{x}$). In this paper, all terms "increasing" and "decreasing" are in the non-strict sense, and inequalities should be interpreted component-wise when applied to vectors.

1. A random vector $\mathbf{P}$ of p-values is PRDS on $\mathcal{N}$ (or, simply, satisfies PRDS) if for any null index $k \in \mathcal{N}$ and increasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}(\mathbf{P} \in A \mid P_k \le x)$ is increasing on $[0, 1]$.

2. A random vector $\mathbf{E}$ of e-values satisfies PRDS if for any null index $k \in \mathcal{N}$ and decreasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}(\mathbf{E} \in A \mid E_k \ge x)$ is decreasing on $[0, \infty)$.

We chose to use the version of PRDS in Finner et al. (2009, Section 4) (also see Barber and Ramdas (2017)) which is weaker than the original one used in Benjamini and Yekutieli (2001), but the difference is relatively minor: "$P_k \le x$" in the former is replaced by "$P_k = x$" in the latter.

If the null p-values (e-values) are mutually independent and independent of the non-null p-values (e-values), then PRDS holds; as such, PRDS is a generalization of independence. Moreover, increasing individual transforms do not affect the PRDS property. Further, the PRDS property is preserved when moving from p-values to e-values using calibrators, or vice versa by inversion. We record this fact below.

Fact 1. The PRDS property of p-values and that of e-values are equivalent in the following sense: If $\mathbf{P}$ is a vector of PRDS p-values, and $E_k = f_k(P_k)$ for any calibrator f_k from (2), then $\mathbf{E}$ is a vector of PRDS e-values. Similarly, if $\mathbf{E}$ is a vector of PRDS e-values, then $\mathbf{P}$ is a vector of PRDS p-values, where $P_k = 1/E_k$. As an important example, PRDS holds if p-values/e-values are built on positively correlated (jointly) Gaussian test statistics $X_1, \dots, X_K$ (themselves PRDS). For instance, this includes p-values obtained from Neyman-Pearson tests, i.e., $P_k = 1 - \Phi(X_k)$ where Φ is the standard Gaussian cdf, and e-values obtained from mixture likelihood ratios, $E_k = \int \exp(\delta X_k - \delta^2/2)\, \mathrm{d}\nu_k(\delta)$, where ν_k is an arbitrary mixing distribution; see Example 3 for instance. The PRDS property of these p-values and e-values directly follows from the invariance of PRDS under monotone transforms.

Following the literature, we will study FDR in both the case of PRDS and that of arbitrary dependence. The input of the Benjamini-Hochberg (BH) procedure includes three ingredients:

(a) K realized p-values $p_1, \dots, p_K$ associated to $H_1, \dots, H_K$, respectively;

(b) an FDR level $\alpha \in (0, 1)$;

(c) (optional) dependence information or assumption on p-values, such as independence, PRDS or no information.

For $k \in \mathcal{K}$, let $p_{(k)}$ be the k-th order statistics of $p_1, \dots, p_K$, from the smallest to the largest. The *(base) BH procedure* $\mathcal{D}(\alpha)$ rejects all hypotheses with the smallest k^* p-values, where

$$k^* = \max\left\{k \in \mathcal{K} : \frac{K p_{(k)}}{k} \le \alpha\right\}, \tag{3}$$

with the convention $\max(\varnothing) = 0$, and accepts the rest. We summarize some known results of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) on the procedure $\mathcal{D}(\alpha)$ below. Throughout, we write

$$\ell_K := \sum_{k=1}^{K} \frac{1}{k} \approx \log K.$$

Theorem 1. *For arbitrary p-values and $\alpha \in (0,1)$, the base BH procedure $\mathcal{D}(\alpha)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\alpha)}}{R_{\mathcal{D}(\alpha)}}\right] \le \frac{\ell_K K_0}{K}\alpha \le \ell_K \alpha.$$

Moreover, if the p-values satisfy PRDS, then

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\alpha)}}{R_{\mathcal{D}(\alpha)}}\right] \le \frac{K_0}{K}\alpha \le \alpha,$$

and the first inequality is an equality if the null p-values are iid uniform on $[0,1]$.

By Theorem 1, if the p-values are PRDS, then the base BH procedure has an FDR of at most α. If the p-values are arbitrarily dependent, then we need to replace α in the base BH procedure by α/ℓ_K, resulting in the Benjamini-Yekutieli (BY) procedure $\mathcal{D}'(\alpha) := \mathcal{D}(\alpha/\ell_K)$, which rejects all hypotheses with the smallest k^* p-values, where

$$k^* = \max\left\{k \in \mathcal{K} : \frac{\ell_K K p_{(k)}}{k} \le \alpha\right\}. \tag{4}$$

The extra factor ℓ_K in the BY procedure reflects the fact that it is harder to justify discoveries without the PRDS assumption, which includes independence as a special case. It is well known that this factor cannot be improved in general.

4 The e-BH procedure

In this section, we design an analog of the BH procedure for e-values, which we call the *e-BH procedure*. Similar to the BH procedure, the input of the e-BH procedure includes three ingredients:

(a) K realized e-values $e_1, \dots, e_K$ associated to $H_1, \dots, H_K$, respectively;

(b) an FDR level $\alpha \in (0,1)$;

(c) (optional) distributional information or assumption on e-values.

While (a) and (b) are the same as those of the BH procedure, (c) is somewhat different: in addition to dependence information, e-BH can also accommodate information of marginal distributions, since e-values have more freedom than p-values in terms of their distributions (the former are only constrained by their expectations under the null).

The e-BH procedure can be described in two simple steps.

1. (Optional) Boost the *raw e-values* in (a) using information in (c).
2. Apply the *base e-BH procedure* to the boosted e-values and level α.

We will first describe the base e-BH procedure in step 2 and then explain how to boost e-values in step 1. If there is no available information in (c), we can simply skip step 1 above, and directly apply the base e-BH procedure to the raw e-values, which always has a valid FDR control at level α.

4.1 The base e-BH procedure

Let $e'_1, \dots, e'_K$ be the boosted e-values obtained from step 1 of the procedure. Keep in mind that they can be chosen as identical to the raw e-values $e_1, \dots, e_K$ if there is no information in (c) or one opts to skip the boosting in step 1.

The base e-BH procedure is applying the BH procedure to $1/e'_1, \dots, 1/e'_K$. To be precise, for $k \in \mathcal{K}$, let $e'_{[k]}$ be the k-th order statistic of $e'_1, \dots, e'_K$, sorted from the largest to the smallest so that $e'_{[1]}$ is the largest boosted e-value. The base e-BH procedure $\mathcal{G}(\alpha) : [0, \infty]^K \to 2^{\mathcal{K}}$ then rejects hypotheses with the largest k^*_e (boosted) e-values, where

$$k^*_e = \max \left\{ k \in \mathcal{K} : \frac{k e'_{[k]}}{K} \geq \frac{1}{\alpha} \right\}. \tag{5}$$

We note that, analogous to the BH procedure, there are many equivalent ways of describing the e-BH procedure, including using an estimate of the false discovery proportion as pioneered by Storey (2002), or the self-consistency viewpoint of Blanchard and Roquain (2008). The alternative formulation in the next proposition will be helpful when showing several results on FDR of the e-BH procedure.

Proposition 1. *Let $e'_1, \dots, e'_k$ be the raw or boosted e-values. Define*

$$R(t) := |\{k \in \mathcal{K} : e'_k \geq t\}| \vee 1 \quad \textit{and} \quad t_\alpha := \inf\{t \in [0, \infty) : tR(t) \geq K/\alpha\}. \tag{6}$$

For each k, the base e-BH procedure applied to $e'_1, \dots, e'_K$ and level α rejects H_k if and only if $e'_k \geq t_\alpha$. Moreover, $t_\alpha R(t_\alpha) = K/\alpha$.

A self-contained simple proof of Proposition 1 is put in Appendix B.

4.2 Boosting e-values

To boost e-values in step 1, there are two types of information we can use: marginal distributional information and joint dependence information. Below we first describe our standard choice of boosting, and then the more general choices which can be freely chosen by the user (Remark 3). Although we generally recommend the standard choice, the flexibility of choosing boosting functions allows us to include the p-based methods (like BH and BY) as special cases.

Let $K/\mathcal{K} := \{K/k : k \in \mathcal{K}\}$, and define a truncation function $T : [0, \infty] \to [0, K]$ by letting $T(x)$ be the largest number in $K/\mathcal{K} \cup \{0\}$ that is no larger than x. In other words,

$$T(x) = \frac{K}{\lceil K/x \rceil} \mathbb{1}_{\{x \geq 1\}} \quad \text{with } T(\infty) = K. \tag{7}$$

Note that T truncates x to take only values in $K/\mathcal{K}$ (or zero). The truncation function T will be frequently used in this paper. The standard boosting method is now described below. For each $k \in \mathcal{K}$, take a *boosting factor* $b_k \geq 1$ (the larger, the better) such that, depending on dependence information,

$$\max_{x \in K/\mathcal{K}} x \mathbb{P}(\alpha b_k E_k \geq x) \leq \alpha \qquad \text{if e-values are PRDS;} \tag{8}$$

$$\mathbb{E}[T(\alpha b_k E_k)] \leq \alpha \qquad \text{otherwise,} \tag{9}$$

where the expectation and the probability are computed under the null distribution of E_k, which is sometimes known. In case of a composite null, (8) and (9) need to hold for all probability measures in H_k, so an additional supremum over H_k must be appended to the left hand side. By Markov's inequality, b_k in (8) is larger or equal to that in (9), leading to a stronger boosting under PRDS. Since $\mathbb{E}[E_k] \leq 1$ for a null e-value E_k, (8) and (9) always hold for $b_k = 1$, making $b_k = 1$ always a safe choice

even if there is no additional (marginal or joint) distributional information. Moreover, the boosting factor b_k is equal to 1 if no marginal information on the distributions of the e-values is available. In either case, define the boosted e-values

$$e'_k = b_k e_k$$

for $k \in \mathcal{K}$. On first reading, the reader may skip the next two remarks.

Remark 2. Note that the left-hand sides of (8) and (9) are increasing in b_k, and hence a practical value of b_k can be obtained by simply trying a few choices of $b_k \geq 1$. It would be ideal to find the largest b_k such that (8) or (9) becomes an equality. This is possible in some cases but it may not be possible if the null distribution of E_k is not continuous or not completely specified; see Examples 2, 3, 5 and 6, where we either get a precise value or an analytical approximation of the best b_k. Under an extra assumption, we have a simple formula (Proposition 3) for the best b_k in case e-values are PRDS.

Remark 3. Instead of using the standard method in (8) and (9) described above, the user can instead choose increasing functions $\phi_1, \dots, \phi_K : [0, \infty] \to [0, \infty]$ such that

$$\max_{x \in K/\mathcal{K}} x\mathbb{P}(\alpha\phi_k(E_k) \geq x) \leq \alpha \qquad \text{if e-values are PRDS;} \tag{10}$$

$$\mathbb{E}[T(\alpha\phi_k(E_k))] \leq \alpha \qquad \text{otherwise,} \tag{11}$$

and define the boosted e-values $e'_k = \phi_k(e_k)$ for $k \in \mathcal{K}$. The choice $\phi_k : x \mapsto b_k x$ corresponds to the standard method explained earlier. Note that if e-values are PRDS, then so are the boosted e-values by any choice of boosting, since PRDS is invariant under monotone transforms.

5 The FDR guarantee of the e-BH procedure

We are now ready to state our main result.

Theorem 2. *The e-BH procedure at level α has FDR at most $K_0\alpha/K$. In particular, the base e-BH procedure applied to arbitrarily dependent raw e-values has FDR at most $K_0\alpha/K$.*

The full proof of Theorem 2 follows from Theorems 3 and 4 in the next few sections, which requires delicate technical treatment. Here, we provide a simple proof for the second statement of Theorem 2 that the base e-BH procedure has the desired FDR guarantee. This simple argument illustrates the advantages of working with e-values, and it applies to any "complaint" e-testing procedures including the base e-BH procedure. Following Blanchard and Roquain (2008), who defined self-consistent p-testing procedures, an e-testing procedure $\mathcal{G}$ is said to be *self-consistent at level* $\alpha \in (0, 1)$ if, denoting by $R_\mathcal{G}$ the number of rejections, every rejected e-value e_k satisfies

$$e_k \geq \frac{K}{\alpha R_\mathcal{G}}.$$

Using $1/p_k$ in place of e_k above, we recover the definition from Blanchard and Roquain (2008) with a linear shape function. Clearly, the base e-BH procedure is self-consistent because of (5); moreover, the base e-BH procedure dominates all other self-consistent e-testing procedures by definition.

Proposition 2. *Any self-consistent e-testing procedure at level α has FDR at most $\alpha K_0/K$ for arbitrary configurations of e-values.*

Proof. Let $\mathcal{G}$ be a self-consistent e-testing procedure. Let $\mathbf{E} = (E_1, \dots, E_K)$ be an arbitrary vector of e-variables fed to the testing procedure $\mathcal{G}$. The FDP of $\mathcal{G}$ satisfies

$$\frac{F_\mathcal{G}}{R_\mathcal{G}} = \frac{|\mathcal{G}(\mathbf{E}) \cap \mathcal{N}|}{R_\mathcal{G} \vee 1} = \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{G}(\mathbf{E})\}}}{R_\mathcal{G} \vee 1} \leq \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{G}(\mathbf{E})\}} \alpha E_k}{K} \leq \sum_{k \in \mathcal{N}} \frac{\alpha E_k}{K}, \tag{12}$$

where the first inequality is due to self-consistency. As $\mathbb{E}[E_k] \leq 1$ for $k \in \mathcal{N}$, we have

$$\mathbb{E}\left[\frac{F_{\mathcal{G}}}{R_{\mathcal{G}}}\right] \leq \sum_{k \in \mathcal{N}} \mathbb{E}\left[\frac{\alpha E_k}{K}\right] \leq \frac{\alpha K_0}{K},$$

thus the desired FDR guarantee. □

The above proof can be alternately viewed as verifying the "dependence condition" in Blanchard and Roquain (2008), which holds even for arbitrarily dependent e-values unlike for p-values.[1]

In contrast, general self-consistent p-testing procedures do not have the nice property in Proposition 2 even if the p-values are independent. Indeed, under a condition slightly weaker than PRDS, Su (2018) proves that a self-consistent p-testing procedure[2] has a weaker FDR guarantee $\alpha(1+\log(1/\alpha)) > \alpha$. Hence, for an FDR guarantee of α, the p-testing procedure needs to be run at level $\alpha' < \alpha$ satisfying

$$\alpha'\left(1 + \log\frac{1}{\alpha'}\right) = \alpha. \tag{13}$$

The FDR control inflation to $\alpha(1 + \log(1/\alpha))$ for arbitrary self-consistent p-testing procedures is quite a contrast to the control at level α for arbitrary self-consistent e-testing procedures, and this fact plays an important role in bandit multiple testing (Jamieson and Jain, 2018; Xu et al., 2021).

5.1 Self-consistent structured, post-selection and grouped e-BH

Without further structure, self-consistent p- or e-testing procedures are dominated by the base BH or e-BH procedure, and hence we do not directly apply them for a stand-alone data set of p-values or e-values. Nevertheless, they become useful in multi-armed bandit problems; see Jamieson and Jain (2018) and our Section 7. Self-consistent procedures are also useful in structured settings. Suppose we would like to insist that the rejected set cannot be an arbitrary subset of $\mathcal{K}$, but must be one amongst a given class of sets $\mathcal{S} \subseteq 2^{\mathcal{K}}$. For example, if the hypotheses are structured as a graph, then we may choose $\mathcal{S}$ to be all connected subgraphs. The appropriate modification of the e-BH procedure ("structured e-BH procedure") would try to

> find any set $S \in \mathcal{S}$ such that every e-value in S is larger than $\frac{K}{\alpha|S|}$.

Importantly, solving for the *largest* S may not be computationally feasible, so one may wish to use an approximation algorithm (like a greedy heuristic that grows S by starting with the largest e-value and adding a few elements at a time to this set) to find *some large* $S \in \mathcal{S}$ such that every e-value in S is larger than $\frac{K}{\alpha|S|}$. Such a procedure would immediately satisfy an FDR guarantee, for any $\mathcal{S}$ and dependence structure, by virtue of being self-consistent. In contrast, the "structured BH procedure" in Ramdas et al. (2019a, Section 8) requires the BY reshaping correction in order to maintain validity.

Another application of self-consistency is in post-processing e-values that have been screened or filtered based on the data. Formally, suppose the original set of hypotheses $\mathcal{K}$ has been shrunk to a data-dependent subset $S \subset \mathcal{K}$ in some arbitrary way based on the e-values. This could be based on any subjective choices of the user (based on side-information, prior scientific knowledge, whatever). The question is: how do we now find a subset of S that guarantees FDR control? The "post-selection e-BH procedure" is simple:

> run the e-BH procedure on S at the amended level $\widetilde{\alpha} := \alpha|S|/K$.

[1] We thank a reviewer for this insightful observation.

[2] Su (2018) used the term *compliance* for self-consistency. We thank a reviewer for pointing out the terminology of self-consistency. Blanchard and Roquain (2008) showed that a self-consistent p-testing procedure that is decreasing in each p-value controls FDR at level α for independent or PRDS p-values.

Then every rejected e-value will be larger than $\frac{|S|}{\widetilde{\alpha} R_{\mathcal{G}}}$, which equals $\frac{K}{\alpha R_{\mathcal{G}}}$. Thus, this procedure controls FDR under arbitrary dependence for any initial selection procedure. This is once more in contrast to the "post-selection BH procedure" in Ramdas et al. (2019a, Section 8), where conditions on both the dependence and the selection are needed for FDR control, devoid of which one must use the BY reshaping correction.

Combining the above two ideas (imposition of structural constraints and screening/filtering) in any order immediately lends a great deal of flexibility to the user to design new procedures. For example, it immediately yields a "focused e-BH" procedure, an extension of the recent focused BH algorithm by Katsevich et al. (2021), whose details are omitted for brevity. In contrast to (Katsevich et al., 2021, Theorem 1), our procedure controls FDR under no assumptions whatsoever on the dependence, structure and filtering.

As a final extension, imagine that we are given a partition of the hypotheses into G prespecified groups. It is easy to define group-level e-values by merging the individual level e-values (e.g., averaging under arbitrary dependence, or multiplying under independence). This immediately yields a group-level e-BH procedure that can be used as an initial filtering step before running the post-selection e-BH procedure. For consistently aggregating discoveries across multiple partitions, an e-filter extension of the p-filter algorithm (Ramdas et al., 2019a) may be derived.

5.2 A few remarks on the e-BH procedure

At this point, we have seen that the base e-BH procedure has the desirable FDR control. Let us make a few remarks on the (full) e-BH procedure to further explain its features and subtleties. Some features have been mentioned briefly before and here they are discussed in comparison with alternatives.

1. Although one can always opt to skip the boosting step and directly apply the base e-BH method to raw e-values, boosting can be quite helpful to enhance detection power. The boosting factor b_k may be substantial (see Examples 2, 3, 5 and 6); a boosting factor between 1.3 and 10 is common in stylized settings.

2. If the e-values are independent (or PRDS), then by Theorem 1, the FDR of the base e-BH procedure is at most α. This result is not of primary interest, but is worthy of note. The main advantage of e-BH is that, unlike the BH procedure, the above FDR guarantee holds under arbitrary dependence, where a conversion to p-values would be suboptimal due to the extra $\approx \log K$ correction factor. This point is relevant in case p-values are derived from e-values, for example via universal inference (Wasserman et al., 2020).

3. The boosting in step 1 of the BH procedure can easily incorporate partial information on the null distributions of the e-values. If we know a set of null possible distributions of E_k, then it suffices to calculate a boosting factor under each distribution, and take their infimum. Moreover, if only the null distributions of some e-values are known, we can simply choose $b_k = 1$ for the ones whose distributions we do not know. Note that the post-selection and structured e-BH procedures also enjoy FDR control after boosting under arbitrary dependence.

4. Since b_k may be different for each k, the boosted e-values may not have the same order as the raw e-values. Hence, the full e-BH procedure may reject a hypothesis with a smaller e-value while accepting one with a larger e-value. This feature is intentional. For instance, an observation of $e_k = 1.999$ carries stronger evidence against the null hypothesis of a uniform distribution on $[0, 2]$, compared to an observation of $e_k = 2$ against a null hypothesis of an exponential distribution with mean 1 or an unknown distribution.

5. E-values typically contain less information than p-values (the flip side of needing less assumptions/structure/knowledge/modeling in order to be constructed), and hence a simplistic comparison with the BH procedure applied to p-values is not particularly insightful. Nevertheless,

if the distributions of null e-values are fully known, then the e-BH procedure performs comparably to the BH and the BY procedures; see Section 6.5 for their connection and Section 7 and Appendix A for simulation results.

6. We note a contrasting feature on the FDR of the BH and e-BH procedures. The BH procedure has a desirable FDR control under dependence assumptions (such as independence and PRDS) and a penalty if such assumption is not available (e.g., (Benjamini and Yekutieli, 2001; Fithian and Lei, 2020)). In contrast, the e-BH procedure has a desirable FDR control with no dependence assumption and a boosting of power is possible if dependence assumptions are imposed. Similarly, a penalty in (13) needs to be applied for self-consistent p-testing procedures, which are dominated by BH, but not for self-consistent e-testing procedures.

7. Suppose we may avail ourselves of only approximate e-values that (under the null) satisfy $\mathbb{E}[E_i] \leq 1+\epsilon_i$, or "asymptotic" e-values, which satisfy $\mathbb{E}[E_i] \leq 1+o(1)$, in the sense that as the amount of data used to calculate the e-value increases in size, then the approximate e-value becomes a valid e-value in the limit. Then the corresponding FDR control in the first case is simply bounded by $\frac{\alpha}{K}(K_0 + \sum_{i\in\mathcal{N}} \epsilon_i)$, which is further bounded by the simpler expression $\alpha(1 + \max_i \epsilon_i)$. In the second case, the FDR bound becomes $\alpha(1 + o(1))K_0/K$. In other words, errors in the e-values (which may be due to slightly violated assumptions, or nuisance parameters, etc.) directly and linearly propagate to errors in the FDR control level. This claim can be verified by simply observing the linearity of expectation used in the last line of the proof of Proposition 2. We know of no existing simple, analogous claim for the BH procedure or other p-testing procedures.

In practice, it may be desirable to assign weights based on prior knowledge of each hypothesis. Here, we describe an analog of the weighted BH procedure (Benjamini and Hochberg, 1997). For arbitrary constant weights $w_1, \ldots, w_K$ summing up to K, a similar boosting scheme is obtained by each replacing α with $w_k\alpha$ in (8) and (9). We can alternatively skip the boosting in step 1 and directly use $w_1E_1, \ldots, w_KE_K$ as the input e-values fed to the base e-BH procedure. In either case, the full e-BH procedure still has the valid FDR control at level α, under arbitrary dependence of the e-values as before. This claim is also justified by Theorems 3 and 4 (see the explanation after Theorem 3).

At this point, most (but not all) messages for the practitioner have been delivered and the casual reader may be warned that the paper intentionally takes a rather theoretical turn, and analyzes the e-BH procedure more carefully, comparing repeatedly to the BH procedure or to the case when e-values are simply calibrated p-values. Proofs of all results beyond this section are put in Appendix B.

6 Theoretical analysis

6.1 A technical lemma

While we have provided a direct and simple proof for FDR control of the base e-BH that does not require any correction under arbitrary dependence, the reader may be curious to understand better why exactly BH has to pay an extra $\approx \log K$ factor, but e-BH, on the contrary, allows for boosting. We provide one technical answer here in the form of a lemma, similar to Ramdas et al. (2019a, Lemma 1) in the setting of p-values. Lemma 1 leads to FDR control of the full e-BH procedure applied to boosted e-values.

Recall that for $k \in \mathcal{K}$, a random vector $\mathbf{X} = (X_1, \ldots, X_K)$ is PRDS on X_k if for any decreasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}(\mathbf{X} \in A \mid X_k \geq x)$ is decreasing on $[0, \infty)$.

In order to prepare for the lemma that follows, note that for any positive constant c, we have that

$$\mathbb{E}[c\mathbb{1}_{\{X_k \geq c\}}] \leq \mathbb{E}[X_k], \tag{14}$$

which is simply a restatement of Markov's inequality. The following lemma addresses the situation where c is a random data-dependent quantity $f(\mathbf{X})$, as in the computation of FDR of the e-BH procedure.

Lemma 1. *Take an arbitrary random vector* $\mathbf{X} = (X_1, \ldots, X_K) : \Omega \to [0,\infty]^K$ *and fix* $k \in \mathcal{K}$. *Let* $f : [0,\infty]^K \to [0,\infty)$ *be a measurable function with range* I_f, *and* $X'_k := \sup\{x \in I_f \cup \{0\} : x \le X_k\}$ *so that* $X'_k \le X_k$ *by construction.*

(i) If X_k *is independent of* $\mathbf{X}^{-k} := (X_j)_{j\neq k}$, *then*

$$\mathbb{E}\left[f(\mathbf{X})\mathbb{1}_{\{X_k \ge f(\mathbf{X})\}} \mid \mathbf{X}^{-k}\right] \le \mathbb{E}[X'_k] \le \mathbb{E}[X_k].$$

(ii) If f *is decreasing and* $\mathbf{X}$ *satisfies PRDS on* X_k, *then*

$$\mathbb{E}\left[f(\mathbf{X})\mathbb{1}_{\{X_k \ge f(\mathbf{X})\}}\right] \le \sup_{x\ge 0} x\mathbb{P}(X'_k \ge x) \le \mathbb{E}[X_k].$$

(iii) For any dependence structure, it always holds that

$$\mathbb{E}\left[f(\mathbf{X})\mathbb{1}_{\{X_k \ge f(\mathbf{X})\}}\right] \le \mathbb{E}[X'_k] \le \mathbb{E}[X_k].$$

In particular, if X_k *is a null e-value, all expectations above are bounded by* 1.

Later we will apply Lemma 1 to $\mathbf{X} = \mathbf{E}'$ and a specific choice of f to show Theorems 3 and 4.

Comparing Lemma 1 to Ramdas et al. (2019a, Lemma 1) for the p-value setting, we observe that the first two statements have a very direct parallel. However the third statements are where the difference appears; the p-value statement requires a correction while the above e-value statement does not. To briefly elaborate, recall (14) and correspondingly note that a null p-variable P_k satisfies for any constant $c \in (0,1]$,

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{P_k \le c\}}}{c}\right] \le 1. \tag{15}$$

While Lemma 1 showed that (14) easily generalized from a constant c to a data-dependent $f(\mathbf{X})$, Ramdas et al. (2019a, Lemma 1) showed that a corresponding statement for (15) is true for p-variables only under a PRDS assumption, but otherwise an extra ℓ_K factor is paid under arbitrary dependence.

We anticipate this lemma to aid with the design and proof of FDR procedures with e-values in other contexts—just as the corresponding p-value lemma already has (Ramdas et al., 2019b)—and we shall see how it is applied to e-BH next.

6.2 FDR of e-BH for arbitrary e-values

We analyze some properties of the base e-BH procedure. Below, the e-values we feed into the base e-BH procedure are the boosted e-values from step 1 of the BH procedure. We will omit the term "boosted" while keeping in mind that the base e-BH can also be directly applied to the raw e-values. Let $(E'_1, \ldots, E'_K)$ be the random vector of (boosted) e-values and $(e'_1, \ldots, e'_K)$ be its realized value. The only property we will use on $E'_1, \ldots, E'_K$ is that they are nonnegative. The next result is a more precise analysis of the FDR of the base e-BH procedure under arbitrary dependence.

Theorem 3. *Applied to arbitrary non-negative random variables* $E'_1, \ldots, E'_K$ *and* $\alpha \in (0,1)$, *the base e-BH procedure* $\mathcal{G}(\alpha)$ *satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] = \frac{\alpha}{K}\sum_{k\in\mathcal{N}} \mathbb{E}\left[t_\alpha \mathbb{1}_{\{E'_k \ge t_\alpha\}}\right] \le \frac{K_0}{K} y_\alpha,$$

where t_α *is given by* (6), *and*

$$y_\alpha = \frac{1}{K_0}\sum_{k\in\mathcal{N}} \mathbb{E}\left[T(\alpha E'_k)\right]. \tag{16}$$

In particular, if $E'_1, \ldots, E'_K$ *are the raw e-values or the boosted e-values via* (9) *or* (11), *then* $y_\alpha \le \alpha$.

Let $(E_1, \dots, E_K)$ be the vector of raw e-values. Theorem 3 shows that for any choice of $b_1 E_1, \dots, b_K E_K$ satisfying (9) or $\phi_1(E_1), \dots, \phi_K(E_K)$ satisfying (11), the FDR of the e-BH procedure is at most α under arbitrary dependence. Moreover, the FDR control of the weighted e-BH follows from

$$\frac{K_0}{K} y_\alpha = \frac{1}{K} \sum_{k \in \mathcal{N}} \mathbb{E}\left[T(\alpha E'_k)\right] \le \frac{1}{K} \sum_{k \in \mathcal{N}} w_k \alpha \le \frac{1}{K} \sum_{k \in \mathcal{K}} w_k \alpha = \alpha.$$

The value of y_α depends on the distribution of the boosted e-values as well as the number K of hypotheses. In case the null distributions of the e-values are known, we would ideally set y_α to α (or close to α) by properly choosing b_k in (9) or ϕ_k in (11).

In case $\mathbb{E}[T(\alpha b_k E_k)]$ is not easy to compute, it might be convenient to use a weaker bound

$$\bar{y}_{\alpha,k}(b_k) := \mathbb{E}[\alpha b_k E_k \mathbb{1}_{\{\alpha b_k E_k \ge 1\}}] \ge \mathbb{E}[T(\alpha b_k E_k)], \tag{17}$$

and set $\bar{y}_{\alpha,k}(b_k) \le \alpha$ (ideally an equality) by choosing $b_k \ge 1$. Such a choice of b_k always guarantees the FDR of the e-BH procedure to be at most α by Theorem 3, since

$$\frac{1}{K_0} \sum_{k \in \mathcal{N}} \mathbb{E}[T(\alpha b_k E_k)] \le \frac{1}{K_0} \sum_{k \in \mathcal{N}} \bar{y}_{\alpha,k}(b_k) \le \alpha.$$

In addition to being easier to compute, an advantage of $\bar{y}_{\alpha,k}(b_k)$ is that it depends purely on the distribution of E_k and not on K, and hence the boosted e-value $b_k E_k$ is ready to use for other experiments involving H_k.

Example 2. We illustrate y_α and $\bar{y}_{\alpha,k}(b_k)$ with a popular class of calibrators in Shafer (2021) and Vovk and Wang (2021). Take $\lambda \in (0,1)$ and assume the raw e-values are given by, for $k \in \mathcal{K}$,

$$E_k = \lambda P_k^{\lambda - 1}, \tag{18}$$

where P_k is a uniform random variable on $[0,1]$ if $k \in \mathcal{N}$. We will consider the boosted e-values $bE_1, \dots, bE_K$ where b is a common boosting factor since the null e-values are identically distributed. In this case,

$$y_\alpha = \mathbb{E}[T(\alpha b E_1)] = \int_0^1 T\left(\alpha b \lambda u^{\lambda - 1}\right) \mathrm{d}u.$$

A formula for $\bar{y}_{\alpha,k}(b)$ is simple:

$$\bar{y}_{\alpha,k}(b) = \alpha b \int_0^{(\alpha b \lambda)^{1/(1-\lambda)}} \lambda u^{\lambda - 1} \, \mathrm{d}u = (\lambda^\lambda \alpha b)^{1/(1-\lambda)}.$$

For instance, if $\lambda = 1/2$, then $\bar{y}_{\alpha,k}(b) = (\alpha b)^2/2$. Setting $\bar{y}_{\alpha,k}(b) = \alpha$ yields $b = (2/\alpha)^{1/2}$. In this example, all e-values are boosted by a multiplier of $(2/\alpha)^{1/2}$, which is substantial; e.g., $b \approx 6.32$ if $\alpha = 0.05$. Finally, note that the same boosting factor b above is also valid whenever the p-values are not exactly uniform but instead stochastically larger than uniform.

Example 3. We consider e-values obtained from the likelihood ratio between two normal distributions with different mean and variance 1 as used in the numerical experiment of Vovk and Wang (2021). Take $\delta > 0$ which represents the difference in the alternative and the null means, and assume the raw e-values are given by, for $k \in \mathcal{K}$,

$$E_k = e^{\delta X_k - \delta^2/2} \tag{19}$$

where X_k is a standard normal random variable if $k \in \mathcal{N}$. Note that each null e-value is a log-normal random variable with parameter $(-\delta^2/2, \delta)$. For the boosted e-values $bE_1, \dots, bE_K$, We have

$$\bar{y}_{\alpha,k}(b) = \alpha b \mathbb{E}[E_k \mathbb{1}_{\{E_k \ge 1/(\alpha b)\}}] = \alpha b \Phi\left(\frac{\delta}{2} + \frac{\log(\alpha b)}{\delta}\right),$$

where Φ is the standard Gaussian cdf. Setting $\bar{y}_{\alpha,k}(b) = \alpha$ yields the equation

$$b\Phi\left(\frac{\delta}{2} + \frac{\log(\alpha b)}{\delta}\right) = 1,$$

which can be easily solved numerically. For instance, if $\delta = 3$ and $\alpha = 0.05$, then $b \approx 1.37$ by solving $\bar{y}_{\alpha,k}(b) = \alpha$, and if $\delta = 4$ and $\alpha = 0.05$, then $b \approx 1.11$. These choices of b are all conservative since $\bar{y}_{\alpha,k}(b)$ is a conservative bound for y_α (but b is only slightly conservative if K is large).

For raw e-values $(E_1, \dots, E_K)$ with unspecified distributions, Theorem 3 gives an upper bound $\alpha K_0/K$ on the FDR of the base e-BH procedure. This upper bound is usually quite loose since y_α in (16) is typically much smaller than α; see e.g., Example 2. This upper bound cannot be improved in general without any additional information.

Example 4 (Sharpness of the upper FDR bound in Theorem 3)**.** Consider the following setup. Let the K_0 null raw e-values be given by $E_k = K/(K_0\alpha)\mathbb{1}_A$, $k \in \mathcal{N}$, where A is an event with $\mathbb{P}(A) = \alpha K_0/K$. Moreover, we set all other e-values to 0, which means that they cannot be rejected. Hence, the false discovery proportion is 1 as soon as there is any rejection. It follows that

$$\mathrm{FDR}_{\mathcal{G}(\alpha)} = \mathbb{P}(R_{\mathcal{G}(\alpha)} > 0) = \mathbb{P}(A) = \alpha K_0/K,$$

which is the upper bound provided by Theorem 3.

6.3 FDR of e-BH for PRDS e-values

In case p-values are independent or positively dependent, the BH procedure has a lower FDR guarantee than the one obtained with arbitrary dependence.

We investigate a similar matter for e-values, aiming for a bound better than that of Theorem 3. This is possible via the result of Lemma 1. To summarize, if e-values are PRDS, the base e-BH procedure has a smaller FDR guarantee, and, consequently, the e-BH procedure allows a more powerful boosting in its step 1.

Theorem 4. *Suppose that the raw e-values are PRDS. Applied to arbitrary non-negative random variables $E'_1, \dots, E'_K$ and $\alpha \in (0,1)$, the base e-BH procedure $\mathcal{G}(\alpha)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] \le \frac{K_0}{K} z_\alpha. \tag{20}$$

where

$$z_\alpha = \frac{1}{K_0} \sum_{k \in \mathcal{N}} \max_{x \in K/\mathcal{K}} x\mathbb{P}(\alpha E'_k \ge x). \tag{21}$$

In particular, if $E'_1, \dots, E'_K$ are the raw e-values or the boosted e-values via (8) or (10), then $z_\alpha \le \alpha$.

Theorem 4 shows that, under the assumption of PRDS, using the boosted e-values $b_1E_1, \dots, b_KE_K$ satisfying (8) or $\phi_1(E_1), \dots, \phi_K(E_K)$ satisfying (10), the FDR of the e-BH procedure is at most α. For $b_k \ge 1$, let

$$z_{\alpha,k}(b_k) = \max_{x \in K/\mathcal{K}} x\mathbb{P}(\alpha b_k E_k \ge x). \tag{22}$$

In step 1 of the e-BH procedure, we need to choose a boosting factor $b_k \ge 1$ such that $z_{\alpha,k}(b_k) \le \alpha$ (ideally, an equality). Under an extra condition, a suitable choice of b_k admits a simple formula.

Proposition 3. *Suppose that E_k is a continuously distributed null e-value. Let $q_{1-\alpha}(E_k)$ be the left $(1-\alpha)$-quantile of E_k. If*

$$t \mapsto t\mathbb{P}(E_k \ge t) \text{ is decreasing on } [q_{1-\alpha}(E_k), \infty), \tag{23}$$

then $b_k := (\alpha q_{1-\alpha}(E_k))^{-1}$ is the largest boosting factor which satisfies (8).

Condition (23) is not uncommon as α is typically small and the value $t\mathbb{P}(E_k \ge t)$ goes to 0 as $t \to \infty$ since E_k has a finite mean. Condition (23) is satisfied by, for instance, the e-values in Example 2 as well as their distributional mixtures. For a continuously distributed E_k, condition (23) can be equivalently expressed as

$$\beta \mapsto \beta q_{1-\beta}(E_k) \text{ is decreasing on } [1-\alpha, 1).$$

Similar to $\bar{y}_{\alpha,k}$ in (17), there is a conservative version of $z_{\alpha,k}$ which does not depend on K, given by

$$\bar{z}_{\alpha,k}(b_k) = \sup_{x \ge 1} x\mathbb{P}(\alpha b_k E_k \ge x). \tag{24}$$

In particular, if (23) holds, then $\bar{z}_{\alpha,k}(b_k) = z_{\alpha,k}(b_k)$ for the choice b_k in Proposition 3.

Example 5. We look at the e-values in Example 2. For $k \in \mathcal{N}$, we have

$$t\mathbb{P}(E_k \ge t) = t(\lambda/t)^{1/(1-\lambda)} = (\lambda t^{-\lambda})^{1/(1-\lambda)},$$

which is a decreasing function in t, and thus (23) holds. In this case, for $b \ge 1$,

$$z_{\alpha,k}(b) = \mathbb{P}(\alpha b E_k \ge 1) = (\lambda \alpha b)^{1/(1-\lambda)},$$

and by Proposition 3, the best choice of b_k is

$$b = (\alpha q_{1-\alpha}(E_k))^{-1} = (\lambda \alpha^{\lambda})^{-1}.$$

Since $\bar{y}_{\alpha,k}(b) = (\lambda^{\lambda}\alpha b)^{1/(1-\lambda)}$ in Example 2, we get $z_{\alpha,k}(b)/\bar{y}_{\alpha,k}(b) = \lambda$. Hence, the FDR is improved by a factor of roughly λ from Theorem 3 (arbitrary dependence) to Theorem 4 (PRDS), which could be substantial if λ is small. For instance, if $\lambda = 1/2$, then $b = 2\alpha^{-1/2}$. For $\alpha = 0.05$, we have $b \approx 8.94$, which should be compared with $b \approx 6.32$ in Example 2 under arbitrary dependence.

Example 6. For the setup of e-values in Example 3, (23) does not always hold. . Nevertheless, for $b \ge 1$, $z_{\alpha,k}(b)$ and $\bar{z}_{\alpha,k}(b)$ in (22) and (24) have simple formulas

$$z_{\alpha,k}(b) = \max_{x \in K/\mathcal{K}} x\, \Phi\left(\frac{\log(\alpha b)}{\delta} - \frac{\log x}{\delta} - \frac{\delta}{2}\right),$$

and

$$\bar{z}_{\alpha,k}(b) = \max_{x \ge 1} x\, \Phi\left(\frac{\log(\alpha b)}{\delta} - \frac{\log x}{\delta} - \frac{\delta}{2}\right).$$

Without specifying K, we use $\bar{z}_{\alpha,k}(b)$, and the value b for $\bar{z}_{\alpha,k}(b) = \alpha$ can be easily computed numerically. For instance, if $\delta = 3$ and $\alpha = 0.05$, then $b \approx 7.88$ (compared with $b \approx 1.37$ in Example 3), and if $\delta = 4$ and $\alpha = 0.05$, then $b \approx 10.31$ (compared with $b \approx 1.11$ in Example 3).

6.4 An optimality result

There is a simple way of generating other e-testing procedures similar to the base e-BH procedure by transforming the e-values. In this section, working under the assumption that no distributional information on e-values is available, we focus on a procedure resulting from applying a common transform to all raw e-values. We obtain an optimality of the base e-BH procedure among all such transforms.

Take a strictly increasing and continuous function $\phi : [0, \infty] \to [0, \infty]$ with $\phi(\infty) = \infty$ and $\phi(0) < 1$. We shall call ϕ an *increasing transform*. We design an e-testing procedure $\mathcal{G}(\phi)$ by rejecting the $k^*_{e,\phi}$ hypotheses with the largest e-values, where

$$k^*_{e,\phi} = \max\left\{k \in \mathcal{K} : \frac{k\phi(e_{[k]})}{K} \ge 1\right\}. \tag{25}$$

It is clear that the choice of $\phi : t \mapsto \alpha t$ corresponds to $\tau_\phi = t_\alpha$ in Section 4, which yields the base e-BH procedure $\mathcal{G}(\alpha)$.

Using Theorem 3 by choosing the boosted e-values as $\phi(E_k)/\alpha$, $k \in \mathcal{K}$, the FDR of $\mathcal{G}(\phi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\phi)}}{R_{\mathcal{G}(\phi)}}\right] \le \frac{K_0}{K} y_\phi, \tag{26}$$

where

$$y_\phi := \frac{1}{K_0} \sum_{k \in \mathcal{N}} \mathbb{E}[T(\phi(E_k))].$$

The special choice $\phi : t \mapsto \alpha t$, corresponding to the e-BH procedure, always guarantees $y_\phi = y_\alpha \le \alpha$, where y_α is defined in (16). For a general choice of ϕ, the FDR of $\mathcal{G}(\phi)$ requires the knowledge of y_ϕ, or at least an upper bound.

In the next result, we show that, if we require an FDR guarantee for arbitrary e-values, then the base e-BH procedure is optimal among the class $\mathcal{G}(\phi)$. This is reassuring for us to use the base e-BH procedure as a canonical candidate for handling arbitrarily dependent e-values without further information.

Theorem 5. *Fix $\alpha \in (0,1)$ and K. For any increasing transform ϕ, if $\mathcal{G}(\phi)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\phi)}}{R_{\mathcal{G}(\phi)}}\right] \le \alpha$$

for arbitrary configurations of e-values, then $\mathcal{G}(\phi) \subseteq \mathcal{G}(\alpha)$.

6.5 Applying the e-BH procedure to p-values

In this section, we compare the BY procedure

$$\mathcal{D}'(\alpha) := \mathcal{D}(\alpha/\ell_K)$$

and the e-BH procedure $\mathcal{G}(\alpha)$, as they both require no assumptions on the dependence structure. Although the two procedures are comparable on the dependence assumption, we remark that e-values generally use less information than p-values, as p-values require full specification of the distributions of the test statistics, whereas e-values only require the information of a known mean. Note that information on the boosting factors b_k, $k \in \mathcal{K}$ is contained in the null distribution of the test statistics.

To properly compare the p- and e-based procedures, we need to calibrate between p-values and e-values using calibrators. With such a calibration, the e-BH procedure gives rise to a general class of p-testing procedures with FDR guarantee for arbitrary p-values, similar to $\mathcal{G}(\phi)$ in Section 6.4, and this class includes the base BH and the BY procedures as special cases. On a related note, calibration to e-values serves as a crucial intermediate step in combining p-values under arbitrary dependence for testing a global null (Vovk et al., 2020, Section 5).

Let $\psi : [0,1] \to [0,\infty]$ be a strictly decreasing and continuous function with $\psi(0) = \infty$. We shall call ψ a *decreasing transform*. We design a p-testing procedure $\mathcal{D}(\psi)$ by rejecting the k^*_ψ hypotheses with the smallest p-values, where

$$k^*_\psi = \max\left\{k \in \mathcal{K} : \psi(p_{(k)}) \ge \frac{K}{k}\right\},$$

with the convention $\max(\varnothing) = 0$. A smaller transform function leads to less power of the testing procedure. A similar procedure using different transforms on individual p-values is analyzed in Appendix C. An important example of $\mathcal{D}(\psi)$ is to choose $\psi : p \to \alpha/p$ for some $\alpha \in (0,1)$. Note that

$$\psi(p_{(k)}) \ge \frac{K}{k} \quad \Longleftrightarrow \quad \frac{K p_{(k)}}{k} \le \alpha.$$

In this case, $\mathcal{D}(\psi)$ is precisely the BH procedure $\mathcal{D}(\alpha)$.

The p-testing procedure $\mathcal{D}(\psi)$ is equal to the step-up procedure of Benjamini and Yekutieli (2001) which rejects $k^* := \max\{k \in \mathcal{K} : p_{(k)} \le \alpha_k\}$ hypotheses with the smallest p-values, where $\alpha_k = \psi^{-1}(K/k)$, $k \in \mathcal{K}$. Such a procedure also appears in Blanchard and Roquain (2008) where $\alpha_k = \alpha\beta(k)/K$ and β is called a shape or reshaping function. Our main purpose here is to compare the p- and e-testing procedures, not to propose new step-up p-testing procedures.

The objects $\phi(p_1), \dots, \phi(p_K)$ can be treated as boosted e-values. Hence, using Theorems 3 and 4, we can easily calculate the FDR of $\mathcal{D}(\psi)$.

Proposition 4. *For arbitrary p-values and a decreasing transform ψ, the testing procedure $\mathcal{D}(\psi)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] \le \frac{K_0}{K} y_\psi,$$

where

$$y_\psi := \psi^{-1}(1) + \sum_{j=1}^{K-1} \frac{K}{j(j+1)} \psi^{-1}(K/j). \tag{27}$$

If the p-values are PRDS, then

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] \le \frac{K_0}{K} z_\psi,$$

where

$$z_\psi := \max_{t \in K/\mathcal{K}} t\psi^{-1}(t). \tag{28}$$

If the decreasing transform ψ satisfies

$$t \mapsto t\psi^{-1}(t) \text{ is decreasing on } [1, \infty), \tag{29}$$

similarly to condition (23), we have $z_\psi = \psi^{-1}(1)$. We can replace $[1, \infty)$ by $[1, K]$ or $K/\mathcal{K}$ in (29); the current condition (29) is slightly stronger but it does not depend on K.

For the specific choice $\psi : p \to \alpha/p$ which gives the BH procedure, we have $\psi^{-1} = \psi$,

$$y_\psi = \alpha + \sum_{j=1}^{K-1} \frac{\alpha}{j+1} = \alpha\ell_K, \text{ and } z_\psi = \max_{t \in K/\mathcal{K}} t\psi^{-1}(t) = \alpha.$$

By Proposition 4, $\mathcal{D}(\psi) = \mathcal{D}(\alpha)$ has FDR guarantee $K_0\ell_K\alpha/K$ for arbitrarily dependent p-values, and FDR guarantee $K_0\alpha/K$ for PRDS p-values. This gives an analytical proof of the FDR guarantee of the BH and BY procedures in Theorem 1.

Let f be a calibrator and $\psi = \alpha f$. We can see that $y_\psi \le \alpha$ by Theorem 3 since $f(p_1), \dots, f(p_K)$ are e-values. Conversely, for any decreasing transform ψ satisfying $y_\psi \le \alpha$ and taking values in $K/\mathcal{K}$ (recall that only values of ψ in $K/\mathcal{K}$ matter), the function $f = \psi/\alpha$ is a calibrator since $y_\psi = \int_0^1 \psi(p)\,\mathrm{d}p = \alpha \int_0^1 f(p)\,\mathrm{d}p$. Therefore, all $D(\psi)$ with $y_\psi \le \alpha$ can be obtained via calibration to e-values.

Remark 4. As we mentioned above, via the relationship

$$\psi^{-1}(K/k) = \alpha\beta(k)/K, \tag{30}$$

one obtains the step-up procedure based on the shape function β. Hence, such procedures can also be equivalently expressed via calibration to e-values. In Blanchard and Roquain (2008), an important condition on β is

$$\beta(k) = \int_0^k x\,\mathrm{d}\nu(x), \ k \in \mathcal{K}, \text{ for some probability measure } \nu \text{ on } (0, \infty). \tag{31}$$

Indeed, if (30) and (31) hold, then by (27),

$$y_\psi = \alpha \sum_{j=1}^{K} \frac{\beta(j) - \beta(j-1)}{j} = \alpha \sum_{j=1}^{K} \int_{j-1}^{j} \frac{x}{j} \, \mathrm{d}\nu(x) \le \alpha \nu((0,\infty)) = \alpha.$$

Therefore, Proposition 4 gives that $D(\psi)$ has FDR level at most α, recovering the result in Blanchard and Roquain (2008). Moreover, we can check that (31) is also sufficient for (29), and thus $D(\psi)$ has FDR level at most $\alpha\beta(K)/K$ under PRDS by Proposition 4.

The inequality $\alpha\mathbb{E}\left[t_\alpha f(t_\alpha)\right] \le K_0 z_\alpha$ in the proof of Theorem 4 is generally not an equality. Hence, the FDR bound provided by Proposition 4 may not be sharp in the case of PRDS p-values. Nevertheless, in the next proposition we shall see that this bound is almost sharp under some extra conditions. From there, we obtain a weak optimality result on the base BH procedure.

Proposition 5. *Fix $\alpha \in (0,1)$ and K. For any decreasing transform ψ, if $\mathcal{D}(\psi)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] \le \alpha$$

for arbitrary configurations of PRDS p-values, then $\psi^{-1}(1) \le \alpha$. Moreover, if ψ satisfies (29), then $\mathcal{D}(\psi) \subseteq \mathcal{D}(\alpha)$.

Comparing the optimality of the e-BH procedure in Theorem 5 with the optimality of the BH procedure in Proposition 5, we note two differences: Theorem 5 is stated for arbitrary e-values whereas Proposition 5 is stated for PRDS p-values; Theorem 5 imposes no assumption on ϕ whereas Proposition 5 requires ψ to satisfy (29).

Different from the case of PRDS p-values, $\mathcal{D}'(\alpha)$ and $\mathcal{D}(\psi)$ may not be strictly comparable for arbitrarily dependent p-values. We discuss this issue in Example 7 below. The general message is that, in contrast to the case of PRDS p-values as shown in Proposition 5, the BY procedure is not necessarily always the best for arbitrarily dependent p-values.

Example 7. We consider the calibrators in (18) by choosing $\psi : p \mapsto \theta\lambda p^{\lambda-1}$ for some $\theta > 0$. For simplicity, we take $\lambda = 1/2$ (a similar procedure was proposed by Sarkar (2008)). As we see from Example 2, $\mathcal{D}(\psi)$ has an FDR guarantee of $y_\psi \le \theta^2/2$. To compare with $\mathcal{D}'(\alpha)$, we choose $\theta = (2\alpha)^{1/2}$, so that both procedures have FDR guarantee of α. Let k^* be the number of hypotheses rejected by $\mathcal{D}'(\alpha)$. Note that a sufficient condition for $k^*_\psi \ge k^*$ is $\psi(p_{k^*}) \ge K/k^*$. If we set

$$\gamma := \frac{k^*\alpha}{K\ell_K} \approx p_{k^*},$$

then, approximately, the above condition is

$$\psi(\gamma) \ge \frac{K}{k^*} \quad \Longleftrightarrow \quad \frac{k^*}{K} \ge \frac{2}{\ell_K}.$$

Note that $\ell_K \sim \log K$, and k^*/K is the proportion of rejection among all hypotheses. Hence, $\mathcal{D}(\psi)$ is more powerful than $\mathcal{D}'(\alpha)$ roughly when the proportion of rejections exceeds $2/\log K$. This conclusion is independent of α.

7 Simulations: multiple testing for ordered hypotheses using multi-armed bandits

Here, we conduct simulation studies for a setting that merges multiple testing with an ordered variant of the multi-armed bandit problem. Appendix A contains additional simulation results for a more

classical non-sequential multiple testing setup involving z-tests. While it is more traditional to present the latter application in a paper, we find the former a more conceptually interesting case study.

Consider a multi-armed bandit with K arms, where pulling arm k produces an iid sample $(X_1^k, X_2^k, \dots)$ from a non-negative random reward X^k. The K null hypotheses are $\mathbb{E}[X^k] \le 1$, $k = 1, \dots, K$, which are nonparametric and similar to Example 1. For interpretation's sake, think of X_j^k as follows: each arm is an investment and X_j^k is the return of the k-th investment on its j-th trial. Specifically, imagine that each investment k starts with unit wealth and the wealth can be multiplied by X_j^k when the arm is pulled the j-th time. Our high-level aim is thus to find profitable investments as quickly as possible; although we do not write down a formal objective, this is not necessary to convey our main points. (One may equally think of testing if the k-th drug has an effect or not.)

Consider a simple setting in which we have used side information or prior knowledge to arrange the arms in a prespecified order and we must collect data on arm k before visiting arm $k+1$; thus previous arms cannot be revisited to obtain more samples. Each arm can be pulled at most n times (the budget) before moving on to the next one.

We describe a natural algorithm to tackle this problem (an algorithm consists of two parts: deciding when to stop pulling an arm and move to the next one, and deciding which nulls to reject at the end). Our algorithm will use a multiple testing procedure $\mathcal{D}$ such as BH or e-BH for both steps: we move onto the next arm if we have exhausted the budget for the current arm, or our multiple testing procedure can reject the null with the available information. To formalize this, we take an e-testing procedure as an example. (For a p-testing procedure, e-values are replaced by p-values.) For $k = 1, \dots, K$:

1. At the end of dealing with arm $k-1$, we can summarize the available information as a vector of e-values $\mathbf{e}_{k-1} := (e_1, \dots, e_{k-1}, 1, \dots, 1)$.

2. Denote by $e_{k,j}$ the e-value obtained after pulling the k-th arm j times. We pull arm k and stop pulling after T_k trials if either

   (a) $|\mathcal{D}(e_1, \dots, e_{k-1}, e_{k,T_k}, 1, \dots, 1)| > |\mathcal{D}(\mathbf{e}_{k-1})|$, i.e., there is at least one new discovery, or

   (b) arm k has been pulled $T_k = n$ times, i.e., the budget is exhausted.

3. Set $e_k := e_{k,T_k}$ and move on to arm $k+1$.

After all K arms are finished, we apply the procedure $\mathcal{D}$ to $\mathbf{e}_K := (e_1, \dots, e_K)$ to determine the rejected hypotheses.

Since we are in a nonparametric setting, the natural choices of e-values and p-values are based on e-processes. More precisely, the e-value $e_{k,j}$ and the p-value $p_{k,j}$ are realized by, respectively,

$$E_{k,j} := \prod_{i=1}^{j} X_i^k \quad \text{and} \quad P_{k,j} := \left(\max_{i=1,\dots,j} E_{k,i} \right)^{-1},$$

for each $k = 1, \dots, K$ and $j = 1, \dots, n$. Also, define $E_{k,0} = P_{k,0} = 1$ for simplicity. It is not hard to see that if H_k is true then $(E_{k,j})_{j=0}^n$ is an e-process. The final e-variables E_k and p-variables P_k are obtained by

$$E_k = E_{k,T_k} \quad \text{and} \quad P_k = P_{k,T_k}, \quad k = 1, \dots, K.$$

The fact that the above stopped e-processes yield valid p-values is obtained by invoking Ville's inequality (Shafer et al., 2011; Howard et al., 2020).

The above algorithm is a natural baseline in the sense that, for $\mathcal{D}$ being BH, e-BH or BY, one cannot obtain more rejections than this algorithm, even if all arms are pulled n times. The algorithm may be seen as an ordered version of that of Jamieson and Jain (2018), but a detailed discussion is not necessary here.

It is important to note that, even assuming data across arms are mutually independent, the produced e-values (or p-values) are dependent in a complicated way; a larger previous realization

leads to a smaller threshold for the next stopping rule. Indeed, we have $T_k < n$ only if Step 2(a) was invoked, meaning that $e_{k,T_k} \geq \frac{K}{|\mathcal{D}(\mathbf{e}_k)|\alpha}$. Hence, some negative dependence exists and the BH procedure does not control FDR theoretically. There are, however, two p-testing procedures that have valid FDR: the BY procedure $\mathcal{D}(\alpha/\ell_K)$ which is designed for arbitrary dependence, and the *self-consistent or compliant BH* (cBH) procedure $\mathcal{D}(\alpha')$, where α' is given in (13); i.e., it satisfies $\alpha'(1-\log(\alpha')) = \alpha$ (e.g., $\alpha' = 0.0087$ for $\alpha = 0.05$). Both the BY and the cBH procedure give a FDR guarantee of α in our experiments, although cBH relies on the independence of data across arms (BY and e-BH do not).[3]

Therefore, we will compare the four procedures e-BH, BH, BY and cBH in our numerical experiments, while keeping in mind that BH does not have a valid FDR in theory, although by design it has more rejections than the other three procedures. We summarize the allowed conditions for a valid FDR guarantee for these four methods in Table 1.

Table 1: Conditions for the validity of the testing algorithm. To allow for dependence among data across arms, we only require for $k \in \mathcal{N}$, $\mathbb{E}[X_j^k|\mathcal{G}] \leq 1$ where $\mathcal{G}$ is the available information before we observe X_j^k (weaker than independence used in our experiments); see Example 1.

	arbitrarily dependent data across arms	arbitrarily dependent stopping rules T_k	FDR guarantee in our experiments
e-BH	YES	YES	valid at level $\alpha K_0/K$
BH	NO	NO	not valid
BY	YES	YES	valid at level $\alpha K_0/K$
cBH	NO	YES	valid at level $\alpha K_0/K$

Below we describe the data generating process used in our experiments (note that the design and the validity of the above procedures do not depend on any knowledge of the data generating process). As mentioned above, arms are naturally ordered such that more promising arms come first. More specifically, arm k is non-null with probability $\theta(K-k+1)/(K+1)$ where $\theta \in [0,1]$ is a parameter. The expected number of non-nulls in this setting is $\theta/2$. Further, let s_k be the strength of signal in each non-null hypothesis, which follows an iid exponential distribution with mean μ. With this setting, some non-nulls may have a very weak signal and they are almost impossible to distinguish from a null, whereas some other non-nulls with a strong signal may be rejected quickly with a few pulls. Conditional on s_k, the data $X_1^k, \ldots, X_n^k$ for arm k are iid following a log-normal distribution

$$X^k = \exp\left(Z^k + s_k \mathbb{1}_{\{k \in \mathcal{K} \setminus \mathcal{N}\}} - 1/2\right)$$

where $Z^1, \ldots, Z^K$ are iid standard normal. Note that the data are independent across arms, and hence the cBH procedure is valid (see Table 1).

In the experiments, we set $\alpha = 0.05$ and $\theta = 1/2$ (on average, 1/4 of all hypotheses are non-null). Results for some values of other parameters are summarized in Table 2.

In all settings of Table 2, we see that the e-BH procedure outperforms both BY and cBH by providing more discoveries and using a smaller number of pulls. The BY and cBH procedures are both quite conservative and they perform similarly, with cBH being better for large K as expected. The numbers of true discoveries of e-BH are slightly lower than those of BH, but, as we explained above, BH does not have a theoretical guarantee of FDR (its realized FDP is not larger than α in our experiments, but this may be due to the fact that the null p-values are not uniform).

[3] We briefly explain the validity of cBH in our setting, which is inspired by observations of Jamieson and Jain (2018). First, define the *unobserved* latent p-values $P'_1, \ldots, P'_K$ by $P'_k = P_{k,n}$, $k \in \mathcal{K}$. Note that $P'_1, \ldots, P'_K$ are independent — since every arm is sampled n times, there is no adaptivity, and thus no dependence. Further, $P_k \geq P'_k$ by definition, and so the output (rejection set) of applying BH to $P_1, \ldots, P_K$ is a subset of that of applying BH to $P'_1, \ldots, P'_K$. Therefore, applying BH to $P_1, \ldots, P_K$ can been as a (randomized) self-consistent procedure applied to the independent latent p-values $P'_1, \ldots, P'_K$, and its FDR control under the correction (13) follows by noting that the proof technique in Su (2018) directly applies to randomly selected self-consistent rejection sets.

Table 2: Simulation results in the multi-armed bandit setting, where R is the number of rejected hypothesis, B is the proportion of unused budget (number of unused pulls divided by nK), and TD is the number of true discoveries (rejected non-null hypotheses). Each number is computed over an average of 500 trials. The default values of parameters are $K = 500$, $n = 50$ and $\mu = 1$, and each panel may have one parameter value different from the default.

(a) Default

	R	B%	TD	FDP%
e-BH	74.4	11.42	73.2	1.58
BH	77.0	11.44	75.3	2.13
BY	70.6	10.06	70.4	0.31
cBH	71.1	10.16	70.8	0.36

(b) $K = 2000$

	R	B%	TD	FDP%
e-BH	297.6	11.39	293.2	1.48
BH	307.8	11.41	301.4	2.07
BY	281.2	9.95	280.4	0.26
cBH	284.5	10.15	283.5	0.36

(c) $n = 10$

	R	B%	TD	FDP%
e-BH	47.7	3.99	47.3	0.83
BH	49.3	4.01	48.7	1.06
BY	38.4	2.77	38.4	0.08
cBH	39.2	2.85	39.2	0.11

(d) $n = 100$

	R	B%	TD	FDP%
e-BH	79.1	13.48	77.9	1.50
BH	81.3	13.50	79.5	2.13
BY	76.4	12.36	76.1	0.35
cBH	76.7	12.44	76.4	0.41

(e) $\mu = 0.5$

	R	B%	TD	FDP%
e-BH	43.5	5.77	42.9	1.54
BH	46.3	5.80	45.3	2.13
BY	39.6	4.66	39.5	0.27
cBH	40.1	4.74	40.0	0.35

(f) $\mu = 2$

	R	B%	TD	FDP%
e-BH	97.4	16.46	95.9	1.54
BH	99.3	16.47	97.2	2.07
BY	94.3	15.23	94.1	0.29
cBH	94.6	15.32	94.3	0.35

For many other settings of parameters and distributions of s_k that we tried, the relative performance of the four methods is always qualitatively similar.

8 A real-data example: detecting cryptocurrencies with positive expected return

In this section we provide a real-data example in the spirit of Example 1 to detect cryptocurrencies (which we call *coins*, like `Bitcoin`, for brevity) with positive expected return. This analysis can be easily conducted for other financial assets or trading strategies.

We collect daily price data of coins from Jan 4, 2015 to Jun 30, 2021 (price data are obtained from `CoinGecko`). We choose to start from 2015 so that we have a few hundreds of coins to work with. There were 496 coins that had trading prices on Jan 4, 2015 (the list is from `CoinMarketCap`), among which 126 were active ("alive") on Jun 30, 2021. We order the coins by their market capitalization in Jan 2015 from the largest to the smallest (e.g., the first being `Bitcoin`).

Let $Y_{k,j}$ be the price of coin k at the end of the j-th month, $j = 1, \dots, T$ and $k = 1, \dots, K$. In our data set, $T = 78$ and $K = 496$ (or 100, 200; see later). Let $X_{k,j} = Y_{k,j}/Y_{k,j-1}$ be the j-th month growth of the coin k, $j \ge 1$, where $Y_{k,0}$ is the initial price of coin k. The k-th null hypothesis is $\mathbb{E}[X_{k,j} \mid \mathcal{F}_{j-1}] \le 1$ for $j = 1, \dots, T$; that is, during the considered period of time, coin k has a monthly expected growth no larger than 1 given previous market information $\mathcal{F}_{j-1}$; see Example 1. We are interested in which coins generate positive expected return during the period. (However, our goal is not to predict future price movements, on which we do not have much to say.)

For each coin k, we start with an initial wealth of \$1, invest $\lambda \in [0,1]$ proportion of our money into the coin and rebalance every month. We ignore transaction costs. The portfolio wealth process $(W_{k,t})_{t=0,1,\dots,T}$ is given by

$$W_{k,0} = 1 \quad \text{and} \quad W_{k,t} = \prod_{j=1}^{t}(1 - \lambda + \lambda X_{k,j}), \quad t = 1, \dots, T.$$

Under the k-th null hypothesis, $(W_{k,t})_{t=0,1,\dots,T}$ is an e-process. Similarly to Section 7, we construct e-values and p-values by

$$E_k := W_{k,T} \quad \text{and} \quad P_k := \left(\max_{t=0,1,\dots,T} W_{k,t} \right)^{-1}.$$

In general, λ is allowed to depend on time and past data, and that would correspond to an adaptive trading strategy; we choose λ to be a constant here for simplicity. Note that E_k above does not require a disclosure of the intermediate wealth values, which may not be available in more complicated applications where privacy or propriety is concerned.

We consider two simple strategies, $\lambda = 1$ which corresponds to buying and holding the coin, and $\lambda = 1/2$ which corresponds to monthly rebalancing to maintain half wealth invested in the coin. Note that for E_k with $\lambda = 1$, we do not need intermediate price data other than the initial and the terminal prices. The vanished coins (dead during the considered period, price data not included in `CoinGecko`) are treated as having insignificant p-values and e-values; recall that for the procedures in this paper, p-values larger than α and e-values less than $1/\alpha$ can be safely treated as 1.

Among the 496 coins listed in Jan 2015, many had tiny market capitalization and did not have meaningful trading activities. For this reason, investing in the full set of 496 coins may not be the most appropriate due to liquidity. Instead, it may be sensible to invest only in 200 or 100 coins with the largest market capitalization in Jan 2015. We thus consider the three cases of $K = 496$ (126 alive in Jun 2021), $K = 200$ (92 alive) and $K = 100$ (54 alive).

The coin returns and the portfolio values are highly dependent in a complicated way. We will apply the e-BH procedure and the BY procedure to the e-values and p-values to select coins, as both methods are valid under arbitrary dependence. The results are reported in Table 3.

Table 3: Number of coins selected by e-BH and BY procedures

	buy and hold			50%-rebalancing			30%-rebalancing		
#total coins	496	200	100	496	200	100	496	200	100
e-BH 5%	3	4	0	42	28	13	21	12	6
BY 5%	11	0	0	15	11	6	15	4	5
e-BH 10%	13	7	2	60	48	24	32	22	8
BY 10%	18	7	0	25	15	7	13	7	7

We observe that the e-BH procedure produces more discoveries than the BY procedure in almost all considered settings, except for $(K, \lambda) = (496, 1)$. In addition, we impose essentially no model assumptions in this application, and the e-BH procedure can be applied even if only initial and final prices are observed. On a side note, for most coins, the 50%-rebalancing strategy produces higher terminal wealth $W_{k,T}$ than the buy-and-hold and 30%-rebalancing strategies, and this could partially be explained by a combination of the very high volatility and the high return of coin prices during the considered period.

9 Conclusion

We have introduced the e-BH procedure to achieve FDR control with e-values. Some highlights of the e-BH procedure are summarized below:

1. It works for arbitrarily dependent e-values (Theorem 2);
2. it requires no information on the configuration of the input e-values, and also works for weighted e-values (Section 5.2);
3. it allows for boosting e-values if partial distributional information is available on some e-values (Section 4.2);
4. it gives rise to a class of p-testing procedures which include both the BH procedure and the BY procedure as special cases along with proofs of FDR control (Proposition 4);
5. it is optimal among the class of e-testing procedures $\mathcal{G}(\phi)$ in the setting of arbitrary e-values (Theorem 5).

Although the e-BH procedure is primarily designed for cases where e-values are directly available (not calibrated from bonafide p-values), even the p-value calibrated e-BH procedure may outperform the BY procedure in some settings (see the simulation results in Appendix A). There are two more complicated situations where the e-BH procedure can be applied:

1. Natural e-values are available for some hypotheses, whereas natural p-values are available for the others. Depending on proportion, one may calibrate them all to p-values or all to e-values.
2. A hypothesis has both an e-value and a p-value (or even multiple p-values and e-values), which may be dependent and obtained from different experiments. In such a situation, the power comparison of e- and p-based procedures depends on the quality of those e-values and p-values. One may choose to combine them into a single e-value or p-value for each hypothesis, e.g., using the methods in Vovk and Wang (2021).

Thorough theoretical and empirical studies of the last two "hybrid" situations are needed to better understand the comparative advantages of p-values and e-values.

Finally, an important issue on the FDR literature is to incorporate certain assumptions on the dependence structure among p-values. A popular choice is to impose a multivariate Gaussian structure; see e.g., Guo et al. (2014), Barber and Candès (2015), Delattre and Roquain (2015) and Fithian and Lei (2020). One may naturally wonder whether there is additional boosting for e-values in a multivariate Gaussian setting, and this leads to interesting questions for future studies.

Acknowledgements

We thank an Editor, an Associate Editor, and two referees for helpful comments on an earlier version of the paper. We also thank Peng Ding, Lihua Lei, Song Mei, and Vladimir Vovk for many useful discussions, and Ziyu Chi for excellent research assistance. AR acknowledges NSF DMS CAREER grant 1945266. RW acknowledges financial support from NSERC grants (RGPIN-2018-03823, RGPAS-2018-522590).

References

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

Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society Series B*, 79(4):1247–1268.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance*, 65(1):179–216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188.

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992.

Delattre, S. and Roquain, E. (2015). New procedures controlling the false discovery proportion via Romano–Wolf's heuristic. *The Annals of Statistics*, 43(3):1141–1177.

Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328):1278–1302.

Embrechts, P., Wang, B., and Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, 19(4):763–790.

Finner, H., Dickhaus, T., and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, 37(2):596–618.

Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence. Technical Report arXiv:2007.10438. To appear in the *Annals of Statistics*.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 32:1035–1061.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101:1408–1417.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26:584–597. Correction: 28:464.

Grünwald, P., de Heide, R., and Koolen, W. M. (2020). Safe testing. Technical Report arXiv:1906.07801.

Guo, W., He, L., and Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics*, 42(3):1070–1101.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Jamieson, K. G. and Jain, L. (2018). A bandit approach to sequential experimental design with false discovery control. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3660–3670.

Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *The Annals of Statistics*, 41(4):2029–2074.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Katsevich, E., Sabatti, C., and Bogomolov, M. (2021). Filtering the rejection set while preserving false discovery rate control. *Journal of the American Statistical Association*.

Ramdas, A., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2019a). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47:2790–2821.

Ramdas, A., Chen, J., Wainwright, M. J., and Jordan, M. I. (2019b). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika*, 106(1):69–86.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2021). Testing exchangeability: fork-convex hulls, supermartingales and e-processes. *International Journal of Approximate Reasoning*.

Sarkar, S. K. (2008). Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference*, 138(4):1072–1084.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Su, W. (2018). The FDR-linking theorem. Technical Report arXiv:1812.08965.

Vovk, V., Wang, B., and Wang, R. (2020). Admissible ways of merging p-values under arbitrary dependence. Technical Report arXiv:2007.14208. To appear in the *Annals of Statistics*.

Vovk, V. and Wang, R. (2019). Confidence and discoveries with e-values. *arXiv preprint arXiv:1912.13292*.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Waudby-Smith, I. and Ramdas, A. (2020a). Confidence sequences for sampling without replacement. *Neural Information Processing Systems*.

Waudby-Smith, I. and Ramdas, A. (2020b). Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Xu, Z., Wang, R., and Ramdas, A. (2021). A unified framework for bandit multiple testing. *Neural Information Processing Systems*.

A Simulations for correlated z-tests

We provide further simulation results in a classic setting, where test statistics X_k, $k \in \mathcal{K}$ are generated from correlated z-tests. The null hypotheses are $\mathrm{N}(0,1)$ and the alternatives are $\mathrm{N}(\delta,1)$, where we take $\delta := -3$ throughout the section. We generate $K - K_0$ observations from the alternative distribution $\mathrm{N}(\delta,1)$ and then K_0 observations from the null distribution $\mathrm{N}(0,1)$.

We take the p-values as

$$P(x) := \Phi(x), \tag{32}$$

where $x \in \{X_k : k \in \mathcal{K}\}$; these are the p-values found using the most powerful test given by the Neyman–Pearson lemma. The raw e-values are, following Vovk and Wang (2021), the likelihood ratios

$$E(x) := \frac{\exp(-(x-\delta)^2/2)}{\exp(-x^2/2)} = \exp(\delta x - \delta^2/2) \tag{33}$$

of the alternative to the null density.

We report the numbers of discoveries for the (base) BH, BY, e-BH (boosted), and base e-BH procedures, in several different settings in Table 4. In particular, we consider two settings of negative correlation (thus, PRDS fails to hold). We set the target FDR level $\alpha \in \{10\%, 5\%, 2\%\}$. All numbers in Table 4 are produced an the average from 1,000 trials.

From the simulation results in Table 4, we can see that the BH procedure has slightly more discoveries than the e-BH procedure for PRDS e-values via (8); recall that we do not expect e-BH to outperform BH when PRDS and precise (standard uniform) p-values are available (cf. Proposition 5). The e-BH procedure for arbitrarily dependent (AD) e-values performs slightly better than the BY procedure in most settings, and this advantage is more pronounced in the case of a large experiment (Table 4b, left), as the BY correction ℓ_K is penalized by a large K. Only in the case of sparse signal and small α (Table 4b, last two columns), BY performs better than e-BH. This is consistent with Example 7 (with different e-values from our experiments), where we see that BY outperforms e-BH if the proportion of rejections is very small.

These simulation results suggest that, even in case p-values are available, the e-BH procedure (with boosted e-values) has competitive performance in most situations, especially when PRDS fails to hold (thus the base BH does not have a correct FDR in theory).

B Omitted proofs from the paper

Proof of Proposition 1. For $t \in [0,\infty)$, let $f(t)$ be the number of true null hypotheses with an e-value e'_k larger than or equal to t. Define the quantity

$$Q(t) = tR(t)/K,$$

and it is clear that

$$t_\alpha = \inf\{t \in [0,\infty) : Q(t) \geq 1/\alpha\}.$$

Clearly $t_\alpha \in [1/\alpha, K/\alpha]$ since $Q(t) \leq t$ and $R(t) \geq 1$. Since Q only has downside jumps and $Q(0) = 0$, we know $Q(t_\alpha) = 1/\alpha$, and thus

$$t_\alpha R(t_\alpha) = \frac{K}{\alpha}. \tag{34}$$

If $e'_k \geq t_\alpha$, then H_k is rejected by the definition of e-BH. If $e'_{[k]} < t_\alpha$, then by definition of Q, we have

$$\frac{k e'_{[k]}}{K} \leq \frac{R(e'_{[k]}) e'_{[k]}}{K} = Q(e'_{[k]}) < 1/\alpha.$$

Thus, each H_k is rejected by $\mathcal{G}(\alpha)$ if and only if $e'_k \geq t_\alpha$. □

Table 4: Simulation results for correlated z-tests, where $\rho_{i,j}$ is the correlation between two test statistics X_i and X_j for $i \neq j$. Each cell gives the number of rejections and, in parentheses, the realized FDP (in %). Each number is computed over an average of 1,000 trials.

(a) Independent and positively correlated tests, $K = 1000$, $K_0 = 800$

	$\rho_{ij} = 0$			$\rho_{ij} = 0.5$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	177.3 (8.01)	148.7 (4.07)	115.0 (1.63)	180.0 (7.00)	144.8 (3.64)	109.8 (1.50)
e-BH PRDS	171.8 (7.07)	147.6 (3.95)	114.6 (1.62)	170.2 (5.71)	142.5 (3.35)	108.0 (1.50)
BY	101.1 (1.10)	78.8 (0.57)	53.2 (0.22)	96.6 (1.03)	76.7 (0.50)	55.0 (0.20)
e-BH AD	109.4 (1.41)	85.4 (0.68)	54.6 (0.24)	103.1 (1.32)	81.4 (0.70)	56.6 (0.28)
base e-BH	97.5 (1.00)	70.6 (0.43)	36.9 (0.11)	91.9 (0.97)	69.1 (0.45)	43.6 (0.16)

(b) Independent tests with large number of hypotheses

	$K = 20,000$, $K_0 = 10,000$			$K = 20,000$, $K_0 = 19,000$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	9567 (5.00)	8564 (2.49)	7164 (1.00)	681.3 (9.58)	520.2 (4.79)	357.7 (1.93)
e-BH PRDS	9092 (3.60)	8330 (2.13)	7124 (0.98)	681.3 (9.58)	509.3 (4.54)	312.1 (1.40)
BY	5956 (0.48)	4818 (0.24)	3417 (0.10)	254.1 (0.89)	177.6 (0.46)	103.1 (0.19)
e-BH AD	6811 (0.80)	5809 (0.44)	4384 (0.18)	271.0 (1.02)	159.5 (0.39)	51.4 (0.07)
base e-BH	6426 (0.64)	5234 (0.31)	3509 (0.10)	224.8 (0.69)	109.2 (0.21)	16.4 (0.01)

(c) Negatively correlated tests, $K = 1000$, $K_0 = 800$.

	$\rho_{ij} = -1/(K-1)$			$\rho_{ij} = -0.5\mathbb{1}_{\{\lvert i-j\rvert=1\}}$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	177.7 (8.14)	149.0 (4.09)	115.2 (1.61)	177.2 (8.10)	148.8 (4.00)	115.3 (1.62)
e-BH PRDS	172.0 (7.13)	147.9 (3.98)	114.9 (1.59)	171.5 (7.13)	147.7 (3.89)	114.9 (1.61)
BY	101.2 (1.08)	78.8 (0.52)	53.3 (0.20)	101.3 (1.11)	78.8 (0.56)	53.2 (0.22)
e-BH AD	109.7 (1.38)	85.5 (0.65)	54.6 (0.22)	109.8 (1.40)	85.6 (0.69)	54.6 (0.24)
base e-BH	97.8 (0.98)	70.7 (0.40)	37.2 (0.11)	97.6 (0.99)	70.7 (0.41)	36.7 (0.12)

Proof of Lemma 1. Note that $f(\mathbf{X}) \leq X'_k$ on the event $\{X_k \geq f(\mathbf{X})\}$.

(i) If X_k is independent of $\mathbf{X}^{-k}$, then

$$\mathbb{E}\left[f(\mathbf{X})\mathbb{1}_{\{X_k \geq f(\mathbf{X})\}} \mid \mathbf{X}^{-k}\right] \leq \mathbb{E}\left[X'_k \mathbb{1}_{\{X_k \geq f(\mathbf{X})\}} \mid \mathbf{X}^{-k}\right] \leq \mathbb{E}\left[X'_k\right] \leq \mathbb{E}[X_k].$$

(ii) Let $g : x \mapsto \mathbb{P}(X'_k \geq x)$ and $P_k := g(X'_k)$. Clearly P_k is a null p-value. For each $y \geq 0$, since $\mathbf{X}$ is PRDS on X_k and $g \circ f$ is an increasing function, the function $x \mapsto \mathbb{P}(g \circ f(\mathbf{X}) \leq y \mid X_k \geq x)$ is decreasing on $[0, \infty)$. Noting that P_k is a decreasing function of X_k, the function $t \mapsto \mathbb{P}(g \circ f(\mathbf{X}) \leq y \mid P_k \leq t)$ is increasing on $[0, 1]$. Using Lemma 1 of Ramdas et al. (2019a), we have

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{P_k \leq g \circ f(\mathbf{X})\}}}{g \circ f(\mathbf{X})}\right] \leq 1.$$

It follows that

$$\begin{aligned}\mathbb{E}[f(\mathbf{X})\mathbb{1}_{\{E_k \geq f(\mathbf{X})\}}] &= \mathbb{E}[f(\mathbf{X})\mathbb{1}_{\{E'_k \geq f(\mathbf{X})\}}] \\ &\leq \mathbb{E}[f(\mathbf{X})\mathbb{1}_{\{P_k \leq g\circ f(\mathbf{X})\}}] \\ &= \mathbb{E}\left[f(\mathbf{X})\mathbb{P}(E'_k \geq f(\mathbf{X}))\frac{\mathbb{1}_{\{P_k \leq g\circ f(\mathbf{X})\}}}{g\circ f(\mathbf{X})}\right] \\ &\leq \sup_{x\geq 0} x\mathbb{P}(X'_k \geq x)\mathbb{E}\left[\frac{\mathbb{1}_{\{P_k \leq g\circ f(\mathbf{X})\}}}{g\circ f(\mathbf{X})}\right] \\ &\leq \sup_{x\geq 0} x\mathbb{P}(X'_k \geq x).\end{aligned}$$

(iii) It follows directly that

$$\mathbb{E}\left[f(\mathbf{X})\mathbb{1}_{\{X_k \geq f(\mathbf{X})\}}\right] \leq \mathbb{E}\left[X'_k\mathbb{1}_{\{X_k \geq f(\mathbf{X})\}}\right] \leq \mathbb{E}[X'_k] \leq \mathbb{E}[X_k].$$

If X_k is a null e-value, then $\mathbb{E}[X_k] \leq 1$ by definition. □

Proof of Theorem 3. By Proposition 1 and (34), we obtain that the FDR of e-BH equals

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] = \mathbb{E}\left[\frac{f(t_\alpha)}{R(t_\alpha)}\right] = \frac{\alpha}{K}\mathbb{E}\left[t_\alpha f(t_\alpha)\right]. \tag{35}$$

We will apply Lemma 1 to $\mathbf{X} = \mathbf{E}'$ and the function f given by $f(\mathbf{X}) = t_\alpha$, where t_α is treated as a function of $\mathbf{E}'$. Note that $I_f = \alpha^{-1}K/\mathcal{K}$ by (34). In this case, $X'_k = T(\alpha X_k)/\alpha$. By Lemma 1 (iii) and (35), we have

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] = \frac{\alpha}{K}\mathbb{E}\left[\sum_{k\in\mathcal{N}} t_\alpha\mathbb{1}_{\{E'_k \geq t_\alpha\}}\right] \leq \frac{1}{K}\mathbb{E}\left[\sum_{k\in\mathcal{N}} X'_k\right] = \frac{K_0}{K}y_\alpha,$$

thus showing the first inequality. If $E'_1, \dots, E'_K$ are the raw e-values or the boosted e-values obtained by (9) or (11), then $y_\alpha \leq \alpha$ by construction. □

Proof of Theoroem 4. Write $\mathbf{E}' = (E'_1, \dots, E'_K)$. For each $k \in \mathcal{N}$, denote by

$$z_{\alpha,k} := \max_{x\in K/\mathcal{K}} x\mathbb{P}(\alpha E'_k \geq x) = \max_{x\geq 1} x\mathbb{P}(T(\alpha E'_k) \geq x) \leq \mathbb{E}[T(\alpha E'_k)].$$

Note that $t_\alpha = f(\mathbf{E}')$ for a decreasing function f with range $I_f = \alpha^{-1}K/\mathcal{K}$. Using the notation in Lemma 1 with $\mathbf{X} = \mathbf{E}'$, we have

$$X'_k = \sup\{x \in I_f \cup \{0\} : x \leq E'_k\} = \frac{T(\alpha E'_k)}{\alpha}.$$

Lemma 1 (ii) gives

$$\mathbb{E}[t_\alpha\mathbb{1}_{\{E'_k \geq t_\alpha\}}] \leq \sup_{x\geq 0} x\mathbb{P}(X'_k \geq x) = \max_{x\geq 1}\frac{x}{\alpha}\mathbb{P}(T(\alpha E'_k) \geq x) = \frac{z_{\alpha,k}}{\alpha}.$$

Hence, by (35),

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] = \frac{\alpha}{K}\mathbb{E}\left[t_\alpha f(t_\alpha)\right] = \frac{\alpha}{K}\sum_{k\in\mathcal{N}}\mathbb{E}[t_\alpha\mathbb{1}_{\{E'_k \geq t_\alpha\}}] \leq \frac{1}{K}\sum_{k\in\mathcal{N}} z_{\alpha,k}.$$

Note that $z_\alpha \leq y_\alpha$ by Markov's inequality. Hence, if $E'_1, \dots, E'_K$ are raw e-values, then $z_\alpha \leq y_\alpha \leq \alpha$. If $E'_1, \dots, E'_K$ are boosted e-values, then $z_\alpha \leq \alpha$ by construction. □

Proof of Proposition 3. Since $b_k = (\alpha q_{1-\alpha}(E_k))^{-1}$, (23) implies

$$t \mapsto t\mathbb{P}(b_k E_k \geq t) \text{ is decreasing on } [1/\alpha, \infty).$$

Therefore, $z_{\alpha,k}(b_k)$ in (22) satisfies $z_{\alpha,k}(b_k) = \mathbb{P}(\alpha b_k E_k \geq 1)$. Direct calculation gives

$$z_{\alpha,k}(b_k) = \mathbb{P}(\alpha E_k \geq 1/b_k) = \mathbb{P}(\alpha E_k \geq \alpha q_{1-\alpha}(E_k)) = \alpha.$$

Since E_k is continuously distributed, b_k cannot be enlarged without violating (8). □

Proof of Theorem 5. Define

$$\tau_\phi = \inf\{\tau \geq 0 : \phi(\tau)R(\tau) \geq K\}.$$

Similarly to Proposition 1, $\mathcal{G}(\phi)$ rejects H_k if and only if $e_k \geq \tau_\phi$. By Theorem 3, the FDR of $\mathcal{G}(\phi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\phi)}}{R_{\mathcal{G}(\phi)}}\right] = \frac{1}{K}\sum_{k\in\mathcal{N}} \mathbb{E}[\phi(\tau_\phi)\mathbb{1}_{\{E_k \geq \tau_\phi\}}]. \tag{36}$$

First, from (25), only the values of $\phi^{-1}(K/k)$ for $k \in \mathcal{K}$ affects the testing procedure $\mathcal{G}(\phi)$. Note that $\mathcal{G}(\alpha)$ uses the linear transform $t \mapsto \alpha t$. Therefore, if $\phi^{-1}(x) \geq x/\alpha$ for all $x \in K/\mathcal{K}$, then $\mathcal{G}(\phi)$ is dominated by the e-BH procedure $\mathcal{G}(\alpha)$, i.e., $\mathcal{G}(\phi) \subseteq \mathcal{G}(\alpha)$.

Next, suppose for the purpose of contradiction that $\phi^{-1}(K/k_0) < K/(k_0\alpha)$ for some $k_0 \in \mathcal{K}$. Write

$$m = \phi^{-1}(K/k_0) \quad \text{and} \quad m' = \max\{m, K/k_0\}.$$

Note that $m \leq m' < K/(k_0\alpha)$. Consider the following setting of e-values: Let $\mathcal{N} = \mathcal{K}$, i.e., all e-values are null. For $k \in \mathcal{K}$, set $E_k = m'\mathbb{1}_{A_k}$ for some set A_k with $\mathbb{P}(A_k) = 1/m'$. We design the sets A_k, $k \in \mathcal{K}$, in the following way: for each $\omega \in \Omega$,

$$|\{k \in \mathcal{K} : \omega \in A_k\}| = k_0,$$

i.e., exactly k_0 events out of $A_1, \dots, A_K$ occur together (the existence of such an arrangement is easy to verify). Let $A = \bigcup_{k\in\mathcal{K}} A_k$, and we have $\mathbb{P}(A) = K/(k_0 m') \in (\alpha, 1]$.

If event A happens, then exactly k_0 of $E_1, \dots, E_K$ take the value m', and the rest take the value 0. Hence, for $\tau > 0$, we have $R(\tau) = \max(k_0\mathbb{1}_{\{\tau \leq m'\}}\mathbb{1}_A, 1)$. If event A happens, then $\tau_\phi \leq m'$ since $\phi(m')R(m') \geq \phi(m)k_0 = K$, and $\tau_\phi \geq m$ since $\phi(\tau_\phi) \geq K/k_0$. Therefore, $\tau_\phi \in [m, m']$ if A happens. It follows that, for each k,

$$\mathbb{E}[\phi(\tau_\phi)\mathbb{1}_{\{E_k \geq \tau_\phi\}}] \geq \mathbb{E}[\phi(m)\mathbb{1}_{\{E_k \geq m'\}}\mathbb{1}_{A_k}] = \mathbb{E}[\phi(m)\mathbb{1}_{A_k}] = \frac{K}{k_0 m'} > \alpha.$$

Using the FDR guarantee α of $\mathcal{G}(\phi)$ and (36), this leads to

$$\alpha \geq \mathbb{E}\left[\frac{F_{\mathcal{G}(\phi)}}{R_{\mathcal{G}(\phi)}}\right] = \frac{1}{K}\sum_{k\in\mathcal{K}} \mathbb{E}[\phi(\tau_\phi)\mathbb{1}_{\{E_k \geq \tau_\phi\}}] > \alpha,$$

a contradiction. Hence, $\mathcal{G}(\phi)$ is dominated by $\mathcal{G}(\alpha)$. □

Proof of Proposition 4. For $k \in \mathcal{N}$, let $\alpha = \mathbb{E}[T(\psi(P))]$ where P is uniform on $[0, 1]$. If $\psi^{-1}(1) = 0$, then $\mathbb{P}(\psi(P_k) \geq 1) = 0$ for $k \in \mathcal{N}$, and thus $\mathcal{D}(\psi)$ will never reject any hypotheses with a non-zero p-value, leading to a zero FDR for both PRDS and arbitrarily dependent p-values. Next, assume $\psi^{-1}(1) > 0$, which implies $\alpha > 0$.

Define

$$E_k = \frac{T(\psi(P_k))}{\alpha}, \qquad k \in \mathcal{K}, \tag{37}$$

which is a null e-value for $k \in \mathcal{N}$. Note that

$$\psi(p_{(k)}) \geq \frac{K}{k} \quad \Longleftrightarrow \quad e_{[k]} \geq \frac{K}{k\alpha}.$$

Hence, $\mathcal{G}(\alpha)$ applied to $(e_1, \dots, e_K)$ is equivalent to $\mathcal{D}(\psi)$ applied to $(p_1, \dots, p_K)$. Note that, using the notation in Sections 6.2 and 6.3,

$$y_\alpha = \frac{1}{K_0} \sum_{k \in \mathcal{N}} \mathbb{E}\left[T(\psi(P_k))\right] \leq \sum_{j=1}^{K} \frac{K}{j} \left(\psi^{-1}\left(\frac{K}{j}\right) - \psi^{-1}\left(\frac{K}{j-1}\right)\right) = y_\psi,$$

and

$$z_\alpha = \max_{x \in K/\mathcal{K}} x\mathbb{P}(\psi(P_k) \geq x) \leq \max_{x \in K/\mathcal{K}} x\psi^{-1}(x) = z_\psi.$$

By Theorems 3 and 4, the testing procedure $\mathcal{D}(\psi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] = \mathbb{E}\left[\frac{F_{\mathcal{G}(\alpha)}}{R_{\mathcal{G}(\alpha)}}\right] \leq \frac{K_0}{K} y_\alpha \leq \frac{K_0}{K} y_\psi,$$

and y_ψ can be replaced by z_ψ if the p-values are PRDS. □

Proof of Proposition 5. Consider the setup where all all p-values are null, and they are identical following a uniform distribution on $[0, 1]$. Note that identical p-values satisfy PRDS. In this case, $F_{\mathcal{D}(\psi)}/R_{\mathcal{D}(\psi)} = 1$ as soon as there is any discovery. If $\psi^{-1}(1) > \alpha$, then

$$\mathbb{P}(\psi(p_{(K)}) \geq 1) = \mathbb{P}(p_{(K)} \leq \psi^{-1}(1)) = \min\{\psi^{-1}(1), 1\} > \alpha.$$

Hence, the probability of having a false discovery is more than α, violating the assumption that $\mathbb{E}[F_{\mathcal{D}(\psi)}/R_{\mathcal{D}(\psi)}] \leq \alpha$. Therefore, $\psi^{-1}(1) \leq \alpha$.

If ψ satisfies (29) and $\psi^{-1}(1) \leq \alpha$, then we have $t\psi^{-1}(t) \leq \alpha$ for all $t \geq 1$. Hence, $\psi(p) \leq \alpha/p$ for $p \in (0, \alpha]$. As a consequence, $\mathcal{D}(\psi) \subseteq \mathcal{D}(\alpha)$. □

C Using e-BH with multiple decreasing transforms

We can generalize the p-testing procedure $\mathcal{D}(\psi)$ in Section 6.5 to the case of multiple decreasing transforms, which we briefly describe here.

Let $\psi_1, \dots, \psi_K$ be decreasing transforms, and write $r_k := \psi_k(p_k)$, $k \in \mathcal{K}$. Design a p-testing procedure $\mathcal{D}(\psi_1, \dots, \psi_K)$ by rejecting the k^* hypotheses with the largest r_k, where

$$k^* = \max\left\{k \in \mathcal{K} : r_{[k]} \geq \frac{K}{k}\right\},$$

with the convention $\max(\varnothing) = 0$. Here, we are not rejecting hypotheses with the smallest p-values (a hypothesis with a smaller p-value may not be rejected before one with a larger p-value), but rather those with the largest values of r_k; thus the procedure is no longer a step-up procedure in Section 6.5. Recall that for a decreasing transform ψ, y_ψ is defined by (27), and z_ψ is defined by (28). The following result is a stronger version of Proposition 4.

Proposition 6. *For arbitrary p-values and decreasing transforms $\psi_1, \dots, \psi_K$, the testing procedure $\mathcal{D}(\psi_1, \dots, \psi_K)$ satisfies*

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi_1,\dots,\psi_K)}}{R_{\mathcal{D}(\psi_1,\dots,\psi_K)}}\right] \leq \frac{1}{K} \sum_{k \in \mathcal{N}} y_{\psi_k}.$$

If the p-values are PRDS, then

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi_1,\dots,\psi_K)}}{R_{\mathcal{D}(\psi_1,\dots,\psi_K)}}\right] \leq \frac{1}{K} \sum_{k \in \mathcal{N}} z_{\psi_k}.$$

Proof. The proof is similar to that of Proposition 4. Define

$$E_k = \frac{T(\psi_k(P_k))}{\beta}, \qquad k \in \mathcal{K},$$

where $\beta = \max_{k\in\mathcal{N}} \mathbb{E}[T(\psi_k(P_k))]$. Clearly, E_k is an e-value for $k \in \mathcal{N}$. Note that

$$r_{[k]} \geq \frac{K}{k} \qquad \Longleftrightarrow \qquad e_{[k]} \geq \frac{K}{k\beta}.$$

Therefore, $\mathcal{G}(\beta)$ applied to $(e_1, \dots, e_K)$ is equal to $\mathcal{D}(\psi_1, \dots, \psi_K)$ applied to $(p_1, \dots, p_K)$. Using Theorems 3 and 4, $\mathcal{D}(\psi_1, \dots, \psi_K)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi_1,\dots,\psi_K)}}{R_{\mathcal{D}(\psi_1,\dots,\psi_K)}}\right] = \mathbb{E}\left[\frac{F_{\mathcal{G}(\beta)}}{R_{\mathcal{G}(\beta)}}\right] \leq \frac{K_0}{K} y_\beta = \frac{1}{K} \sum_{k\in\mathcal{N}} y_{\psi_k},$$

and y_{ψ_k} can be replaced by z_{ψ_k} if the p-values are PRDS. □