Boosting e-BH via conditional calibration

Junu Lee[1] and Zhimei Ren[1]

[1]Department of Statistics and Data Science, University of Pennsylvania

April 15, 2024

Abstract

The e-BH procedure is an e-value-based multiple testing procedure that provably controls the false discovery rate (FDR) under any dependence structure between the e-values. Despite this appealing theoretical FDR control guarantee, the e-BH procedure often suffers from low power in practice. In this paper, we propose a general framework that boosts the power of e-BH without sacrificing its FDR control under arbitrary dependence. This is achieved by the technique of conditional calibration, where we take as input the e-values and calibrate them to be a set of "boosted e-values" that are guaranteed to be no less—and are often more—powerful than the original ones. Our general framework is explicitly instantiated in three classes of multiple testing problems: (1) testing under parametric models, (2) conditional independence testing under the model-X setting, and (3) model-free conformalized selection. Extensive numerical experiments show that our proposed method significantly improves the power of e-BH while continuing to control the FDR. We also demonstrate the effectiveness of our method through application to the task of identifying individuals with positive treatment effects in an observational study.

1 Introduction

We study the problem of testing m hypotheses simultaneously while controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995). For this task, classical methods (e.g., Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001); Storey (2002)) associate each hypothesis with a p-value and decide on which subset of hypotheses to reject based on these p-values. Recently, the notion of e-values has been proposed for quantifying evidence against the null hypothesis in place of p-values (Shafer et al., 2011; Grünwald et al., 2024; Vovk and Wang, 2021; Grünwald, 2023). To be concrete, an *e-value* for a null hypothesis H_0 is the realization of an *e-variable* E, which obeys

$$E \geq 0 \text{ and } \mathbb{E}_{H_0}[E] \leq 1.$$

In contrast, we recall that a *p-value* for H_0 is the realization of a *p-variable* P, such that

$$\mathbb{P}_{H_0}(P \leq t) \leq t, \text{ for all } t \in (0,1).$$

In what follows, we will not distinguish between e-values (resp. p-values) and e-variables (resp. p-variables) when the context is clear. Per their definitions, both e-values and p-values are summaries of evidence against H_0, where we reject H_0 for small p-values or large e-values. Compared with the p-values, several properties of e-values make them attractive for hypothesis testing. For example, the e-value allows the experimenter to adaptively decide whether to collect new evidence or to stop the experiment; it is also handy for combining evidence from multiple sources (see Section 2 for more discussion).

When it comes to multiple testing, Wang and Ramdas (2022) propose a simple and elegant e-value-based procedure, called the e-BH procedure, that provably controls the FDR at the target level α under unknown arbitrary dependence among the e-values. This is a rather surprising result, since for the p-value-based procedures, the FDR control is only guaranteed under special dependency strucures — e.g., when the p-values are independent or positively correlated — unless one is willing to tolerate an inflated FDR level. Despite this theoretical appeal, the e-BH procedure is observed to be conservative in practice, often achieving an FDR much lower than the target level, which greatly hinders its wide application. It is thus of great interest to improve the power of e-BH without sacrificing the FDR control guarantee.

1.1 A peek at our contribution

In this paper, we propose a general framework that boosts the power of e-BH for a wide class of multiple testing problems when partial information on the dependence structure is accessible. To set the stage, consider m null hypotheses $H_1, H_2, \ldots, H_m$, where the subset of null hypotheses which are true is denoted by $\mathcal{H}_0$. For each $j \in [m] := \{1, 2, \ldots, m\}$, H_j is associated with an e-value e_j. Applying the e-BH procedure to $\boldsymbol{e} := \{e_1, e_2, \ldots, e_m\}$, one obtains a rejection set $\mathcal{R}(\boldsymbol{e}) \subseteq [m]$. The set $\mathcal{R}(\boldsymbol{e})$ is guaranteed to control the FDR, whose formal definition is as follows:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\{j \in \mathcal{R}(\boldsymbol{e})\}}{|\mathcal{R}(\boldsymbol{e})| \vee 1}\right],$$

where $|\cdot|$ denotes the cardinality of a set and $a \vee b = \max(a, b)$ for any $a, b \in \mathbb{R}$. To identify the source of conservativeness in e-BH, we follow Fithian and Lei (2022) and decompose the FDR into the "contribution" of each null hypothesis:

$$\mathrm{FDR} = \sum_{j\in\mathcal{H}_0} \mathrm{FDR}_j := \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{j \in \mathcal{R}(\boldsymbol{e})\}}{|\mathcal{R}(\boldsymbol{e})| \vee 1}\right].$$

Intuitively, if a multiple testing procedure is tight—that is, it uses up all its FDR bugdet—then FDR_j should be close to α/m (or other budget b_j that adds up to α given additional information). As we shall see later, FDR_j's are often much smaller than α/m for e-BH, leading to a loss of power.

At a high level, our proposal is to boost the e-BH procedure by filling these gaps, which requires identifying a sufficient statistic S_j for each $j \in [m]$, conditioning on which we can evaluate the distribution of FDR_j under the null. Operationally, our proposed method takes as input the e-values $\boldsymbol{e}$ and returns a set of *boosted e-values* $\boldsymbol{e}^{\mathfrak{b}} := \{e_1^{\mathfrak{b}}, e_2^{\mathfrak{b}}, \ldots, e_m^{\mathfrak{b}}\}$ that are at least as powerful as the original e-values (theoretically and practically); it then applies the e-BH procedure to $\boldsymbol{e}^{\mathfrak{b}}$ to obtain a rejection set $\mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$, which improves upon $\mathcal{R}(\boldsymbol{e})$ in terms of power, while maintaining the FDR control guarantees.

Our second major contribution is to explicitly instantiate our general framework in three classes of multiple testing problems: (1) testing under a class of parametric models, (2) conditional independence testing in the model-X setting, and (3) model-free comformalized selection. For each problem, we identify the sufficient statistics and provide a concrete way to boost the e-values. As a preview, Figure 1 shows the empirical power improvement of our proposed method over the e-BH procedure for a selection of experiments; through all of this, it continues to theoretically and empirically control the false discovery rate (FDR) at a preset level, In the simulation studies of Sections 4, 5, and 6, we find that power improvement occurs over all settings, not just the specific ones chosen here.

Organization of the paper. We introduce the background on e-values and discuss related literature in Section 2. Our main framework for boosting e-values is presented in Section 3. Sections 4, 5 and 6 contain an instantiation of our methods in testing under parametric models, conditional independence testing in the model-X setting, and model-free conformalized selection, respectively. These three sections are rather stand-alone—readers who are interested in a specific problem can jump directly to the corresponding section after reading the opening sections. Section 7 presents the results from real data analysis. We conclude the paper with a discussion in Section 8.

2 Background

2.1 E-values

The notion of e-values includes many commonly used statistics, such as betting scores, Bayes factors, likelihood ratios, and stopped martingales (e.g., Wasserman et al. (2020); Shafer (2021); Howard et al. (2021); Grünwald (2023); Waudby-Smith and Ramdas (2024)). As mentioned earlier, we reject the null hypothesis H_0 when the e-value is large: for any $\alpha \in (0, 1)$, rejecting H_0 when $e \geq 1/\alpha$ yields a level α test as a consequence of Markov's inequality:

$$\mathbb{P}_{H_0}(e \geq 1/\alpha) = \alpha \cdot \mathbb{E}_{H_0}[e] \leq \alpha.$$

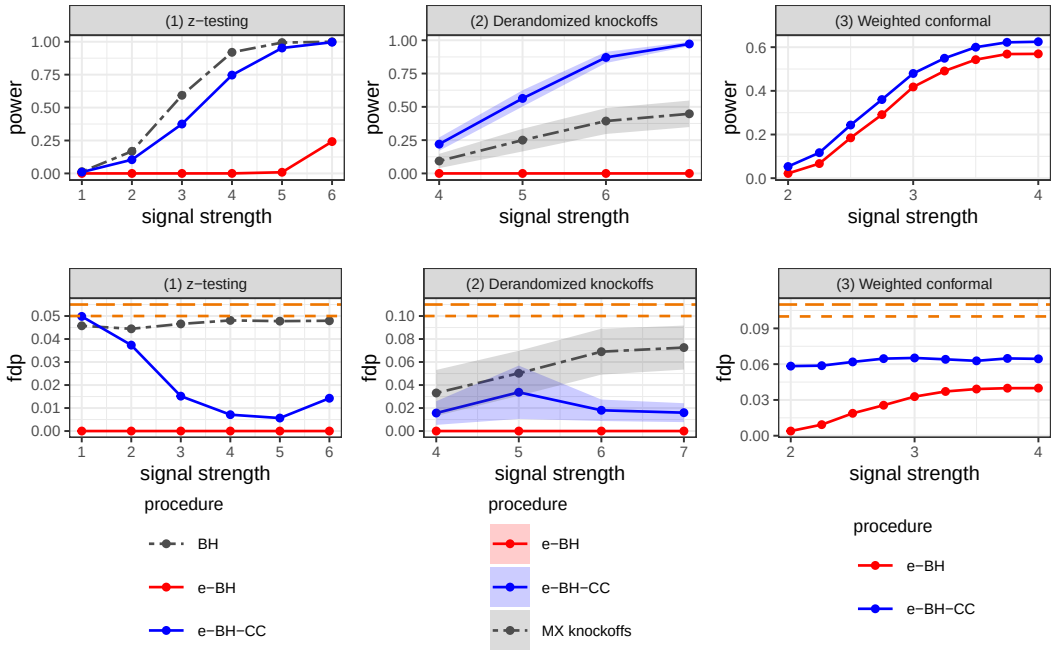

Figure 1: A selection of simulation results from each of the three instantiations of our proposed method, e-BH-CC. The three problem instances are (1) one-sided z-testing using likelihood ratio e-values; (2) conditional independence testing using derandomized knockoffs e-values; and (3) weighted conformal outlier detection using conformal e-values. The plots above show average power and false discovery proportion (FDP) curves over 1000, 100, and 1000 replications, respectively. The dashed gray lines correspond to the relevant baselines (if they exist), and the target FDR level is illustrated by the short-dashed orange line. Shading represents error bars, when they are deemed necessary.

There is also a close connection between the e-value and the p-value: let e be an e-value for a null hypothesis H_0, then $p = 1/e$ is a p-value for H_0 by Markov's inequality,[1] since

$$\mathbb{P}_{H_0}(p \le t) = \mathbb{P}_{H_0}(e \ge 1/t) \le t \cdot \mathbb{E}_{H_0}[e] \le t, \text{ for any } t \in (0,1).$$

Conversely, a p-value p for H_0 can be transformed into an e-value by a "calibrator" (Shafer et al., 2011), defined as a decreasing function $f : [0,1] \mapsto [0,\infty)$, such that

$$\int_0^1 f(t)\, dt = 1.$$

For example, we can take $f(t) = \lambda t^{\lambda-1}$ for some $\lambda \in (0,1)$ to be the calibrator (Shafer, 2021; Vovk and Wang, 2021).

Prior works (Shafer, 2021; Grünwald et al., 2024; Wang and Ramdas, 2022; Grünwald, 2023; Ramdas et al., 2023) have provided fruitful discussions on when one would prefer e-values over p-values, and we refer the readers to them for a comprehensive review. Here, to build intuition and help motivate our examples, we describe two scenarios where e-values are particularly useful.

Imagine a scientist is conducting experiments to test a hypothesis H_0. After collecting the data and performing the data analysis, she obtains an e-value e_1 that fails to reject H_0; having seen e_1, she decides to collect the next batch of data and obtain another e-value e_2. The scientist can then combine the evidence from e_1 and e_2 by taking $e = e_1 e_2$, which is a valid e-value as long as $\mathbb{E}_{H_0}[e_2 \,|\, e_1] \le 1$. In contrast, it is not clear how to combine p-values efficiently in such a sequential experiment setting.

Consider another example where two labs are interested testing the same hypothesis H_0, and their data can be dependent in unknown ways (there can be overlapping cohort in the two studies). Each lab obtains an e-value, e_1 and e_2. To combine the evidence from both labs, one can simply take $e = \frac{1}{2}(e_1 + e_2)$, which is still a valid e-value since $\mathbb{E}_{H_0}[e] = \frac{1}{2}(\mathbb{E}_{H_0}[e_1] + \mathbb{E}_{H_0}[e_2]) \le 1$. The same argument, however, does not hold for p-values.

[1] We use the term "p-value" a bit loosely since we allow the p-value to be larger than 1. One can always transform p into a strict p-value by taking $\min(p, 1)$.

2.2 The e-BH procedure

For m null hypotheses $H_1, H_2, \ldots, H_m$ and their associated e-values $e_1, e_2, \ldots, e_m$, let $e_{(1)}, e_{(2)}, \ldots, e_{(m)}$ denote the ordered e-values in an descending order. The e-BH procedure (Wang and Ramdas, 2022) at level α rejects the hypotheses corresponding to the k^* largest e-values, where

$$k^* = \max\left\{k \in [m] : e_{(k)} \geq \frac{m}{\alpha k}\right\},$$

with the convention $\max \varnothing = 0$. To understand why e-BH is conservative, it is helpful to go through the proof of its FDR control. Let $\mathcal{R}^{\text{e-BH}}$ denote the e-BH rejection set. It can be checked that $j \in \mathcal{R}^{\text{e-BH}}$ if and only if $e_j \geq \frac{m}{\alpha|\mathcal{R}^{\text{e-BH}}|}$, so we can write the FDR of e-BH as

$$\text{FDR} = \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{e_j \geq \frac{m}{\alpha|\mathcal{R}^{\text{e-BH}}|}\}}{|\mathcal{R}^{\text{e-BH}}| \vee 1}\right] \overset{\text{(i)}}{\leq} \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{e_j \frac{\alpha|\mathcal{R}^{\text{e-BH}}|}{m}}{|\mathcal{R}^{\text{e-BH}}| \vee 1}\right] \leq \frac{\alpha}{m}\sum_{j\in\mathcal{H}_0}\mathbb{E}[e_j] \overset{\text{(ii)}}{\leq} \alpha, \tag{1}$$

where step (i) follows from the deterministic inequality $\mathbb{1}\{X \geq t\} \leq X/t$ for any $t > 0$, and step (ii) is due to the definition of e-values. Note that the inequality in (i) is tight if and only if $e_j \in \{0, \frac{m}{\alpha|\mathcal{R}^{\text{e-BH}}|}\}$, which is typically not the case. Step (i) therefore constitutes a major source of the gap between the realized FDR of e-BH and the target level α. The main idea of our proposal, to be introduced in Section 3, is to identify and close this gap by leveraging the distribution of e-values conditional on certain sufficient statistics.

It is worth mentioning that Wang and Ramdas (2022) also provide a method for boosting the e-values by leveraging available information on their *marginal* distributions.[2] Specifically, the method transforms an e-value e_j into $e'_j = b_j e_j$ for some boosting factor $b_j \geq 1$, and then applies the e-BH procedure to $e'_j s$. The boosting factor b_j is chosen to be the largest $b \geq 1$ such that $\mathbb{E}\big[T(\alpha b e_j)\big] \leq \alpha$, where $T(x)$ is the largest element in $\{1, m/(m-1), \ldots, m/2, m\}$ that is no greater than x. For example, when $e_j = \frac{1}{2}p_j^{-1/2}$, with p_j being a uniform random variable on $[0,1]$, one can take $b_j = (2/\alpha)^{1/2}$. Since this boosting scheme mainly leverages the marginal distribution of e-values, we refer to it as the *marginal boosting* scheme. As is often observed in practice, the marginal boosting scheme is not sufficient for closing the gap caused by step (i) in Equation (1).

Finally, we note that for the purpose of FDR control, it suffices to require that the sum of expectation of null e-values is bounded by m (step (ii) in (1)). We call such e-values *generalized e-values*, whose formal definition is given below.

Definition 1 (Generalized e-values)**.** *The non-negative random variables $e_1, e_2, \ldots, e_m$ are called generalized e-values if $\sum_{j\in\mathcal{H}_0}\mathbb{E}[e_j] \leq m$.*

Applying the e-BH procedure to generalized e-values yields a level-α FDR control (Wang and Ramdas, 2022), and this can easily be seen by noting that step (ii) in Equation (1) still holds for generalized e-values. We will see examples of such generalized e-values later.

2.3 Related work

There has been a growing literature on the topic of e-values, including the interpretation and properties of e-values (Shafer et al., 2011; Shafer, 2021; Vovk and Wang, 2021, 2023; Grünwald, 2023), the existence and construction of powerful e-values (Wasserman et al., 2020; Grünwald et al., 2024; Zhang et al., 2023; Larsson et al., 2024), and the use of e-values in various statistical problems (Howard et al., 2020, 2021; Waudby-Smith and Ramdas, 2024; Vladimir and Wang, 2024; Wang et al., 2022; Waudby-Smith et al., 2022). When it comes to multiple testing, as mentioned earlier, Wang and Ramdas (2022) propose an e-value-based procedure for testing multiple hypotheses which controls the FDR under arbitrary dependence. Ignatiadis et al. (2023) study the multiple testing problem when both p-values and e-values are available. Xu and Ramdas (2023b) consider using the e-values for multiple testing in the online setting, while Xu et al. (2022) concerns themselves with building post-selection confidence intervals by inverting the e-values.

It is worth noting that two recent papers (Ramdas and Manole, 2023; Xu and Ramdas, 2023a) discuss the use of external randomness for boosting the power of e-values, where the former focuses on testing

[2] Wang and Ramdas (2022) also provides a more powerful boosting scheme when the p-values are positively dependent on a subset (PRDS). Since the PRDS condition already ensures the FDR control of the BH procedure, in this paper we choose to focus primarily on the more general cases beyond the PRDS condition.

a single hypothesis and the latter considers multiple testing. For example, one could replace e_j with $\widetilde{e}_j := e_j/U_j$, for $U_j \sim \text{Unif}([0,1])$ that is independent of everything else, and then apply the e-BH procedure to the $\widetilde{e}_j$'s, still guaranteeing FDR control. Since $U_j < 1$, such an approach indeed improves the power. But the external randomness can potentially hinder the reproducibility of the results (the procedure can be quite sensitive to the realization of the U_j's and different runs of the algorithm can yield different selections), and could encourage "hacking" the data to obtain a desired result (practitioners can repeatedly sample the U_j's until getting significant results). Our proposed method, in contrast, is deterministic in principle (assuming sufficient computational resources) and is effectively stable across different runs of the algorithm in practice.

There is a substantial line of works on multiple testing with FDR control. This problem is first studied by Benjamini and Hochberg (1995), in which they also propose the Benjamini-Hochberg (BH) procedure that operates on p-values. The BH procedure is only known to control the FDR with independent or positively correlated p-values. Otherwise, a severe correction is needed to ensure the FDR control (Benjamini and Yekutieli, 2001). Subsequent works have investigated the asymptotic FDR control (Genovese and Wasserman, 2004; Storey et al., 2004; Ferreira and Zwinderman, 2006; Farcomeni, 2007). The recent work of Chi et al. (2022) discusses the FDR control of the BH procedure under negative dependence, providing better correction factors than in the arbitrary dependence case. Sarkar (2023) develops a variant of the BH procedure the controls the FDR when testing multivariate normal means against two-sided alternatives. Our work draws inspiration from Fithian and Lei (2022). The authors proposes the dBH procedure that uses conditional calibration to modify the p-value threshold in the BH procedure, thereby achieving finite-sample FDR control, whereas we use conditional calibration to boost the power of the e-BH procedure. We will provide a detailed discussion on the connnection between their method and ours in Section 3.

Beyond the p-value-based mulitple testing procedure, Barber and Candès (2015); Candès et al. (2018) propose the knockoff procedure that controls the FDR by adding "knockoff" variables to the regression. It is shown that the knockoff-based methods also have e-value interpretations (Ren and Barber, 2024); based on this observation, we apply our framework to improving the power of the knockoff method in Section 5.

3 Boosting e-BH using conditional calibration

Given a collection of e-values $\boldsymbol{e} = (e_1, \ldots, e_m)$ corresponding to null hypotheses $H_1, \ldots, H_m$, our method returns a new collection of e-values $\boldsymbol{e}^{\mathfrak{b}} = (e_1^{\mathfrak{b}}, \ldots, e_m^{\mathfrak{b}})$ through a technique called *conditional calibration* (CC) (Fithian and Lei, 2022). Originally proposed as a way to produce separately-calibrated thresholds for *p-values* to return an FDR-controlling rejection set, conditional calibration plays a different role in our procedure—instead of calibrating thresholds for rejection, we boost e-values to more powerful versions of themselves while retaining e-value validity. Thus, we can run e-BH on these boosted e-values to attain a rejection set which still controls the FDR at the pre-specified level $\alpha \in (0,1)$. Furthermore, the rejection set from the boosted method dominates that of e-BH in terms of power.

Suppose we can identify for each $j \in [m]$ a sufficient statistic S_j such that we know the conditional joint distribution $\boldsymbol{e} \mid S_j$ under the null hypothesis H_j. Denote $\mathcal{R}(\boldsymbol{e})$ as the rejection set returned by the e-BH procedure on $\boldsymbol{e}$ at level $\alpha \in (0,1)$. For each $j \in [m]$, define $\widehat{\mathcal{R}}_j(\boldsymbol{e}) := \mathcal{R}(\boldsymbol{e}) \cup \{j\}$ and subsequently define the function

$$\phi_j(c; S_j) := \mathbb{E}\left[\frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{c\widetilde{e}_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}\right\}}{|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|} - \widetilde{e}_j \,\middle|\, S_j\right] \tag{2}$$

where $\widetilde{\boldsymbol{e}} = (\widetilde{e}_1, \ldots, \widetilde{e}_m)$ follows the conditional distribution $\boldsymbol{e} \mid S_j$. Noting that $\phi_j(c; S_j)$ is monotonically non-decreasing in c, we can define the associated critical value

$$\widehat{c}_j := \sup\{c \colon \phi_j(c; S_j) \leq 0\}. \tag{3}$$

Since the function $\phi_j(c; S_j)$ is not necessarily continuous in c, it is possible that $\phi_j(\widehat{c}_j; S_j) > 0$. We then construct our new collection of e-values slightly differently depending on the value of $\phi_j(\widehat{c}_j; S_j)$:

$$e_j^{\mathfrak{b}} = \begin{cases} \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \cdot \mathbb{1}\left\{\widehat{c}_j e_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} & \text{if } \phi_j(\widehat{c}_j; S_j) \leq 0, \\ \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \cdot \mathbb{1}\left\{\widehat{c}_j e_j > \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} & \text{if } \phi_j(\widehat{c}_j; S_j) > 0. \end{cases} \tag{4}$$

We formalize in Theorem 1 that the boosted e-values $(e_1^\flat, \ldots, e_m^\flat)$ are valid e-values, and provide its proof in Appendix A.1.

Theorem 1. *When $(e_1, \ldots, e_m)$ are (resp. generalized) e-values, the boosted e-value $\boldsymbol{e}^\flat = (e_1^\flat, \ldots, e_m^\flat)$ defined in (4) are (resp. generalized) e-values.*

As a consequence of Theorem 1, the rejection set returned from running e-BH on $\boldsymbol{e}^\flat$ at level α satisfies FDR control, i.e., $\mathrm{FDR}[\mathcal{R}(\boldsymbol{e}^\flat)] \leq \alpha$. We summarize this procedure, henceforth referred to as e-BH-CC, in Algorithm 1.

Remark 1. *Since $(e_1^\flat, \ldots, e_m^\flat)$ are valid (generalized) e-values, applying e-BH on them at any level $\alpha_{\text{e-BH}} \in (0, 1)$ yields a rejection set with FDR controlled by $\alpha_{\text{e-BH}}$. In the most general version of e-BH-CC, we denote the level used in constructing the boosted e-values (the α hyperparameter in (2) and (4)) as α_{CC}, and allow α_{CC} to differ from $\alpha_{\text{e-BH}}$. As we shall see shortly, the power improvement of e-BH-CC is theoretically guaranteed when $\alpha_{\text{CC}} = \alpha_{\text{e-BH}}$. In what follows, we will assume that $\alpha_{\text{CC}} = \alpha_{\text{e-BH}} = \alpha$ unless otherwise specified.*

Algorithm 1: e-BH-CC

Input: e-values $e_1, e_2, \ldots, e_m$; sufficient statistics $S_1, S_2, \ldots, S_m$; target FDR level α.

for $j \in [m]$ **do**
  1. Compute the boosting factor $\widehat{c}_j$.
  2. Construct the boosted e-values $e_j^\flat$ according to (4).

end

Apply e-BH to $(e_1^\flat, e_2^\flat, \ldots, e_m^\flat)$ at level α and obtain the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^\flat)$.

Output: the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^\flat)$.

The reader might wonder why we describe e-BH-CC as "boosting" e-BH. This is because by constructing boosted versions of the e-values and applying e-BH to them, we have created a more powerful procedure than regular e-BH. Specifically, $\mathcal{R}(e^\flat)$ can be interpreted as a rejection set returned by a selection procedure which improves upon regular e-BH by tightening its FDR control. Notably, the process described above is *deterministic* with respect to the original e-values collected—there is no randomness introduced.

3.1 Power improvement

We previously claimed that the boosted rejection set $\mathcal{R}(\boldsymbol{e}^\flat)$ has greater power than $\mathcal{R}(\boldsymbol{e})$. We formalize the claim as follows:

Theorem 2. *Given e-values $\boldsymbol{e} = (e_1, \ldots, e_m)$, denote $\boldsymbol{e}^\flat = (e_1^\flat, \ldots, e_m^\flat)$ to be the boosted e-values from conditional calibration defined in (4). Then $\mathcal{R}(\boldsymbol{e}^\flat) \supseteq \mathcal{R}(\boldsymbol{e})$, where each rejection set comes from running the e-BH procedure at the same level $\alpha \in (0, 1)$.*

Proof of Theorem 2. The claim mainly follows from the fact that $\phi_j(1; S_j)$ for each j. To see why, we invoke the inequality that $\mathbb{1}\{X \geq t\} \leq X/t$ for $t > 0$ on the indicator inside the function ϕ_j:

$$\mathbb{1}\left\{\widetilde{e}_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}\right\} \leq \frac{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}{m} \cdot \widetilde{e}_j.$$

Therefore,

$$\phi_j(1; S_j) \leq \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|} \cdot \frac{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}{m} \cdot \widetilde{e}_j - \widetilde{e}_j \,\middle|\, S_j\right] = 0.$$

As a result, when $\phi_j(\widehat{c}_j; S_j) \leq 0$, we have $\widehat{c}_j \geq 1$; when $\phi_j(\widehat{c}_j; S_j) > 0$, we have $\widehat{c}_j > 1$.

We will now show that $j \in \mathcal{R}(\boldsymbol{e}) \implies j \in \mathcal{R}(\boldsymbol{e}^\flat)$. The case when $\mathcal{R}(\boldsymbol{e})$ is empty is trivial, so we assume otherwise. For each $j \in \mathcal{R}(\boldsymbol{e})$, the containment $\widehat{\mathcal{R}}_j(\boldsymbol{e}) \supseteq \mathcal{R}(\boldsymbol{e})$ actually attains set equality, so

$$e_j^\flat = \begin{cases} \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|} \cdot \mathbb{1}\left\{\widehat{c}_j e_j \geq \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|}\right\} & \text{if } \phi_j(\widehat{c}_j; S_j) \leq 0, \\ \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|} \cdot \mathbb{1}\left\{\widehat{c}_j e_j > \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|}\right\} & \text{if } \phi_j(\widehat{c}_j; S_j) > 0. \end{cases}$$

Recall that in the case of $\phi_j(\widehat{c}_j; S_j) \leq 0$, $\widehat{c}_j \geq 1$ and in the case of $\phi_j(\widehat{c}_j; S_j) > 0$, $\widehat{c}_j > 1$. In either case, the indicator in the above crystallizes to 1 since $e_j \geq \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|}$ by virtue of j having been selected

by e-BH. Therefore, $e_j^{\mathfrak{b}} = \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|}$ for each $j \in \mathcal{R}(\boldsymbol{e})$. Since there are $|\mathcal{R}(\boldsymbol{e})|$ many such indices, the e-BH procedure will select the set $\{e_j^{\mathfrak{b}} : j \in \mathcal{R}(\boldsymbol{e})\}$ at the very least, proving the claim. □

3.2 Implementing e-BH-CC

To calculate $\widehat{c}_j$ in (3), we assume access to the conditional expectation ϕ_j through knowledge of the distribution $\boldsymbol{e} \mid S_j$ under H_j. In practice, it is unreasonable to expect an analytical form of ϕ_j for anything but the simplest of settings. In that sense, the e-BH-CC procedure we previously outlined amounts to an oracle algorithm.

Therefore, to design an implementation of e-BH-CC that is much more practical, we will avoid analytically calculating $\phi_j(\cdot; S_j)$ in favor of numerically evaluating it by using i.i.d. resamples from $\boldsymbol{e} \mid S_j$. With these resamples, we can use Monte-Carlo estimation in order to evaluate $\phi_j(\cdot; S_j)$ at any point. However, this introduces a tradeoff between computational cost and accuracy. Considering that, naïvely, we desire to estimate the critical value of the function, it is not immediately clear how to translate the ability to resample into an efficient e-BH-CC implementation whose FDR control does not severely and unpredictably suffer from Monte-Carlo error.

In this subsection, we outline our computationally efficient Monte-Carlo method for impelementing e-BH-CC, which only requires resamples from $\boldsymbol{e} \mid S_j$. In our implementation, we forego estimating the critical value $\widehat{c}_j$ and rather evaluate $\mathbb{1}\{\widehat{c}_j e_j \geq m/(\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|)\}$ directly, since this determines the value of $e_j^{\mathfrak{b}}$. We show that it suffices to evaluate $\phi_j(\cdot; S_j)$ at a specific value, which we can Monte-Carlo estimate using our resamples. By making this simplification, we crucially avoid the issue of finding $\widehat{c}_j$. Our implementation also uses anytime-valid methods to control the MC estimation error in an online manner, which has methodological and computational benefits to be seen later. This approach takes inspiration from the work of Luo et al. (2022), which similarly uses conditional calibration to improve the power of the knockoff filter (Barber and Candès, 2015; Candès et al., 2018). We draw a distinction with their method, as ours applies in a general multiple testing framework—boosting the power of the knockoff filter is a specific application of e-BH-CC, as detailed in Section 5.

3.2.1 Evaluating the numerator

As aluded to previously, the only unknown quantity of $e_j^{\mathfrak{b}}$ is the numerator (the denominator is directly a function of the original e-values). To approach evaluating the numerator, we observe the following equivalences of events:

$$\begin{aligned}
&\text{when } \phi_j(\widehat{c}_j; S_j) \leq 0: \quad \left\{\widehat{c}_j e_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \iff \left\{\widehat{c}_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j\right\} \iff \{\phi_j(\widetilde{c}_j; S_j) \leq 0\}, \\
&\text{when } \phi_j(\widehat{c}_j; S_j) > 0: \quad \left\{\widehat{c}_j e_j > \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \iff \left\{\widehat{c}_j > \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j\right\} \iff \{\phi_j(\widetilde{c}_j; S_j) \leq 0\},
\end{aligned} \tag{5}$$

where $\widetilde{c}_j = \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j$. The last event $\{\phi_j(\widetilde{c}_j; S_j) \leq 0\}$ may confuse the astute reader, who will recall that $\phi_j(c; S_j)$ is an expression with c inside the conditional expectation. By evaluating $\phi_j(\cdot; S_j)$ at the random $\widetilde{c}_j$, we mean to evaluate it at the *value taken* by $\widetilde{c}_j$, rather than substituting c with $\widetilde{c}_j$ in (2). This is equivalent to evaluating

$$\mathbb{E}\left[\frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{\left(\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j\right)\widetilde{e}_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}\right\}}{|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|} - \widetilde{e}_j \,\middle|\, S_j, \boldsymbol{e}\right] \tag{6}$$

with the expectation over the distribution $\widetilde{\boldsymbol{e}} \mid S_j, \boldsymbol{e}$. In (6), we can write the resample $\widetilde{\boldsymbol{e}} = f(S_j, U_j)$ for some function f and a uniform ramdom variable U_j that is independent of everything else. From this representation, we can clearly see that $\widetilde{\boldsymbol{e}} \perp\!\!\!\perp \boldsymbol{e}_j \mid S_j$, and the conditional expectation (6) collapses to only conditioning on S_j. Equations (5) and (6) imply that we can evaluate the numerator by estimating the conditional mean (given S_j and $\boldsymbol{e}$) of

$$\frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{\widetilde{e}_j/e_j \geq |\widehat{\mathcal{R}}_j(\boldsymbol{e})|/|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|\right\}}{|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|}.$$

Assume we have K i.i.d. resamples from $\boldsymbol{e} \mid S_j$. From each resample $\widetilde{\boldsymbol{e}}^{(k)}, k \in [K]$, we can compute

$$E_{j,k} = \frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{\widetilde{e}_j^{(k)}/e_j \geq |\widehat{\mathcal{R}}_j(\boldsymbol{e})|/|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}}^{(k)})|\right\}}{|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}}^{(k)})|} \tag{7}$$

thereby giving us K samples

$$E_{j,1}, E_{j,2}, \ldots, E_{j,K} \overset{\text{i.i.d.}}{\sim} \frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{\widetilde{e}_j/e_j \geq |\widehat{\mathcal{R}}_j(\boldsymbol{e})|/|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|\right\}}{|\widehat{\mathcal{R}}_j(\widetilde{\boldsymbol{e}})|} \mid S_j.$$

The typical Monte-Carlo estimator for $\phi_j(\widetilde{c}_j; S_i)$ is the average of the samples $\overline{E}_{j,K} \coloneqq \frac{1}{K}\sum_{k=1}^{K} E_{j,k}$. When K is large, giving us arbitrarily precise Monte-Carlo estimation, we can replace $\mathbb{1}\{\phi_j(\widetilde{c}_j; S_j) \leq 0\}$ with $\mathbb{1}\left\{\overline{E}_{j,K} \leq 0\right\}$ in the construction of $e_j^{\mathfrak{b}}$ with no repercussions.

When K is instead chosen with regard to computational budget restrictions, then we may experience Monte-Carlo estimation error. The goal is then to control the effect of such error on the resulting FDR of the overall procedure. Importantly, we are more preoccupied with having confidence in the *sign* of $\phi_j(\widetilde{c}_j; S_j)$ rather than its value. Therefore, we can use confidence intervals to control the error from estimating the sign, which filters through as an additive penalty to the resulting FDR.

For some fixed $K \in \mathbb{Z}^+$ and for each j, produce K i.i.d. resamples $\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(K)}$ conditional on S_j, and compute $E_{j,1}, \ldots, E_{j,K}$ according to (7). Define $\alpha_{\text{CI}} = \alpha_0|\mathcal{R}^{\text{e-BH}}(\boldsymbol{e})|/m$, for some $\alpha_0 \in (0,1)$ corresponding to the Monte-Carlo error budget. Using the observations $E_{j,1}, \ldots, E_{j,K}$, construct a $(1-\alpha_{\text{CI}})$-coverage confidence interval $C_{j,K} \coloneqq C_{j,K}(E_{j,1}, \ldots, E_{j,K})$ such that

$$\mathbb{P}\big(\phi(\widetilde{c}; S_j) \in C_{j,K} \mid S_j, \boldsymbol{e}\big) \geq 1 - \alpha_{\text{CI}}.$$

Let $U_{j,K}$ be the upper endpoint of $C_{j,K}$. Define the CI-approximated boosted e-values

$$e_j^{\mathfrak{b},\text{CI}} = m \cdot \frac{\mathbb{1}\{U_{j,K} \leq 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}. \tag{8}$$

We summarize the above procedure in Algorithm 2, and formalize the validity of the CI-approximated boosted e-values in Proposition 1.

Algorithm 2: e-BH-CC with CI-approximated boosted e-values

Input: e-values $e_1, e_2, \ldots, e_m$; sufficient statistics $S_1, S_2, \ldots, S_m$; target FDR level α; number of Monte-Carlo samples K; Monte-Carlo error budget α_0.

$\alpha_{\text{CI}} \leftarrow \alpha_0|\mathcal{R}^{\text{e-BH}}(\boldsymbol{e})|/m$.
for $j \in [m]$ **do**
  Generate $\widetilde{\boldsymbol{e}}^{(1)}, \ldots, \widetilde{\boldsymbol{e}}^{(K)} \overset{\text{i.i.d.}}{\sim} \boldsymbol{e} \mid S_j$.
  Compute $E_{j,1}, \ldots, E_{j,K}$ according to (7).
  Construct a $(1-\alpha_{\text{CI}})$ confidence interval $C_{j,K}$ for $\phi_j(\widetilde{c}_j; S_j)$.
  Construct the CI-approximated boosted e-values as in (8).
end
Apply e-BH to $(\boldsymbol{e}_1^{\mathfrak{b},\text{CI}}, \ldots, \boldsymbol{e}_m^{\mathfrak{b},\text{CI}})$ at level α and obtain the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{CI}})$.

Output: the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{CI}})$.

Proposition 1. *Suppose $\boldsymbol{e} = (e_1, \ldots, e_m)$ are generalized e-values. Running e-BH on the collection of CI-approximated boosted e-values $\boldsymbol{e}^{\mathfrak{b},\text{CI}} = \{e_j^{\mathfrak{b},\text{CI}}\}_{j\in[m]}$ defined in (8) with the target FDR level α and the Monte-Carlo error budget α_0, we have*

$$\text{FDR}\big[\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{CI}})\big] \leq \alpha + \alpha_0.$$

If we replace the $(1-\alpha_{\text{CI}})$ confidence interval with an asymptotic $(1-\alpha_{\text{CI}})$ confidence interval, we obtain instead asymptotic FDR control, as formalized by the following corollary.

Corollary 1. *Replacing the $(1-\alpha_{\text{CI}})$ confidence interval in Proposition 1 with an asymptotic $(1-\alpha_{\text{CI}})$ confidence interval leads to the analogous conclusion*

$$\lim_{K\to\infty} \text{FDR}\big[\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{CI}})\big] \leq \alpha + \alpha_0,$$

where K is the number of resamples from $\boldsymbol{e} \mid S_j$.

The proof of Proposition 1, as well as that of Corollary 1, can be found in Appendix A.2.

3.2.2 Online Monte-Carlo estimation with error control

When K is fixed ahead of time, the user may experience a disappointing event: the Monte-Carlo estimate $\overline{E}_n$ is negative, yet the confidence interval contains zero, making the resulting boosted e-value zero as well. Given this, it would be quite tempting to continue resampling from $\boldsymbol{e} \mid S_j$, hopefully until the confidence interval is finally contained within $\mathbb{R}^{\leq 0}$. Unfortunately, this adaptive mechanism will break the error control of the confidence interval, which has downstream implications for FDR control.

To address this issue in general, we resort to the *anytime-valid confidence sequence* (AVCS) (see e.g. Ramdas et al. (2023) for a review). A $(1-\alpha)$-coverage AVCS for some parameter θ is a sequence of confidence intervals $\{(L_k, U_k)\}_{k\geq 1}$ such that

$$\mathbb{P}\big(\forall k \in \mathbb{N}, \theta \in [L_k, U_k]\big) \geq 1-\alpha.$$

This is in contrast to the common confidence interval, where the "for all k" qualifier is outside of the probability. Usually, the AVCS is used as an online version of a confidence interval: at each time step k, a new sample X_k is added to the existing sequence of samples to construct the latest iterate of the confidence interval $[L_k, U_k]$ in a way such that the miscoverage probability of this process at any point is at most α.

Using the AVCS, we can ignore K and replace the confidence interval in Proposition 1 with its anytime-valid variant: for each $j \in [m]$, accrue samples $E_{j,1}, E_{j,2}, \ldots \mid S_j$ by resampling $\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)}, \ldots \mid S_j$ and using (7). Define $\alpha_{\text{AVCS}} = \alpha_0 |\mathcal{R}^{\text{e-BH}}(\boldsymbol{e})|/m$ for some α_0 corresponding to the Monte-Carlo error budget. We then construct a $(1-\alpha_{\text{AVCS}})$-coverage anytime-valid confidence sequence $\{C_{j,k} \coloneqq C_{j,k}(E_{j,1}, \ldots, E_{j,k})\}_{k\geq 1}$, and define the AVCS-approximated boosted e-values

$$e_j^{\mathfrak{b},\text{AVCS}} = m \cdot \frac{\mathbb{1}\{\exists k \in \mathbb{N} \colon U_{j,k} \leq 0\}}{\alpha |\widehat{\mathcal{R}}_j(\boldsymbol{e})|},$$

where $U_{j,k}$ is the upper endpoint of $C_{j,k}$. Algorithm 3 summarizes the procedure of implementing e-BH-CC with AVCS-approximated boosted e-values, and Proposition 2 formalizes the FDR control of the resulting rejection set.

Algorithm 3: e-BH-CC with ACVS-approximated boosted e-values

Input: e-values $e_1, e_2, \ldots, e_m$; sufficient statistics $S_1, S_2, \ldots, S_m$; target FDR level α; Monte-Carlo error budget α_0.

$\alpha_{\text{CI}} \leftarrow \alpha_0 |\mathcal{R}^{\text{e-BH}}(\boldsymbol{e})|/m$.
for $j \in [m]$ **do**
  Initialization: $k \leftarrow 0$ and $C_{j,0} = \mathbb{R}$.
  while $U_{j,k} > 0$ and $L_{j,k} \leq 0$ **do**
    $k \leftarrow k+1$.
    Generate $\widetilde{\boldsymbol{e}}^{(k)} \sim \boldsymbol{e} \mid S_j$.
    Compute $E_{j,k}$ according to (7).
    Construct the kth interval $C_{j,k}$ in the $(1-\alpha_{\text{AVCS}})$ confidence sequence for $\phi_j(\widetilde{c}_j; S_j)$.
  end
  Construct the AVCS-approximated boosted e-values $e_j^{\mathfrak{b},\text{AVCS}} = m \cdot \frac{\mathbb{1}\{U_{j,k} \leq 0\}}{\alpha |\widehat{\mathcal{R}}_j(\boldsymbol{e})|}$.
end
Apply e-BH to $(\boldsymbol{e}_1^{\mathfrak{b},\text{AVCS}}, \ldots, \boldsymbol{e}_m^{\mathfrak{b},\text{AVCS}})$ and obtain the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{AVCS}})$.

Output: the rejection set $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{AVCS}})$.

Proposition 2. *Suppose $\boldsymbol{e} = (e_1, \ldots, e_m)$ are generalized e-values. Running e-BH on the collection of AVCS-approximated boosted e-values $\boldsymbol{e}^{\mathfrak{b},\text{AVCS}} = \{e_j^{\mathfrak{b},\text{AVCS}}\}_{j\in[m]}$ with the target FDR level α and the target Monte-Carlo error budget α_0, we have*

$$\text{FDR}\big[\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b},\text{AVCS}})\big] \leq \alpha + \alpha_0.$$

In practice, we can replace the $(1-\alpha_{\text{AVCS}})$-AVCS by an asymptotic $(1-\alpha_{\text{AVCS}})$-coverage AVCS (Waudby-Smith et al., 2021) in Proposition 2 when k is large enough. We delegate the definition of an asymptotic AVCS and the proof of Proposition 2 in Appendix A.3.

3.2.3 Filtering the (potential) rejection set

One additional improvement, in both computational feasibility and power, is to restrict the focus onto potentially rejectable hypotheses. Since we have to repeat the process of resampling and Monte-Carlo estimation for each $j \in [m]$, we can save an entire iteration of computation by choosing to not boost specific e-values.

For example, one rudimentary filter is to avoid the indices $\{j : e_j = 0\}$. For all such j, the boosted e-value $e_j^{\mathfrak{b}}$ will be zero directly by construction (4) regardless of the crystallization of the critical value $\widehat{c}_j$. Sections 5 and 6 contain two examples of problem settings where the corresponding e-values can be zero with positive probability. More generally, any e-value that is designed as an "all-or-nothing bet" (Shafer, 2021) takes value $1/p$ with probability $p > 0$ and 0 otherwise.

Another possible strategy to filter out low-potential discoveries is to use preliminary test statistics, such as p-values or linear model coefficients. Take, for example, the m-dimensional z-testing problem setting (covered in-depth in Section 4.1). Each hypothesis $H_j : \mu_j \leq 0$ has a corresponding p-value p_j and e-value e_j. To control the FDR at α, the BH procedure rejects p_j if it lies below $\alpha\widehat{k}/m$ for some data-driven $\widehat{k} \in [m]$; therefore, a prerequisite to rejecting p_j is that it is at most α. We can define a filter set $M := \{j : p_j \leq \alpha\}$ and use it to construct *masked* boosted e-values:

$$e_j^{\mathfrak{b},M} := e_j^{\mathfrak{b}} \mathbb{1}\{j \in M\}. \tag{9}$$

Since $\mathbb{E}[e_j^{\mathfrak{b},M}] = \mathbb{E}[e_j^{\mathfrak{b}} \mathbb{1}\{j \in M\}] \leq \mathbb{E}[e_j^{\mathfrak{b}}]$, the masked e-values are also valid. To see examples of filters used in simulations, we direct the reader to the details of Sections 4.3 and 5.4.

Using (9) for any filter $M \supseteq \mathcal{R}(\boldsymbol{e})$ gives us a collection of zeroed-out e-values, over which we can run e-BH and control the FDR. At this point, the reader may express concern that filtering may cause e-BH-CC to become overly conservative. We argue this is not the case, in the sense that $M \cap \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}}) = \mathcal{R}(\boldsymbol{e}^{\mathfrak{b},M})$ exactly. To illustrate this more specifically, assume that there is exactly one index j such that $j \notin M$ and $M \cup \{j\} = \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$. Then all other $k \in \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}}), k \neq j$ will still end up in $\mathcal{R}(\boldsymbol{e}^{\mathfrak{b},M})$. That is, the "accidental" filtering out of $e_j^{\mathfrak{b}}$ will not interfere with the fate of any e-value kept by the filter M. This is a straightforward observation stemming from the fact that the e-BH rejection threshold over $\boldsymbol{e}^{\mathfrak{b}}$, denoted $\widehat{t}_{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b}})$, has the following characterization:

$$\widehat{t}_{\text{e-BH}}(\boldsymbol{e}^{\mathfrak{b}}) \begin{cases} = \frac{m}{\alpha|\mathcal{R}(\boldsymbol{e})|} & \text{when } \mathcal{R}(\boldsymbol{e}) = \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}}) \\ \leq \frac{m}{\alpha(|\mathcal{R}(\boldsymbol{e})|+1)} & \text{when } \mathcal{R}(\boldsymbol{e}) \subsetneq \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}}). \end{cases}$$

Combining this with the fact that boosted e-values take magnitude at least $\frac{m}{\alpha(|\mathcal{R}(\boldsymbol{e})|+1)}$, we conclude that $\mathcal{R}(\boldsymbol{e}^{\mathfrak{b}}) \cap M = \mathcal{R}(\boldsymbol{e}^{\mathfrak{b},M})$

We also previously claimed that filtering improves the power of our method. First, we note that M is generally designed to contain $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e})$ in order to preserve the uniform power improvement. In addition, when implementing e-BH-CC in practice with Monte-Carlo estimation, we use a confidence interval or AVCS and follow Proposition 1 or 2, respectively. In both cases, the confidence level (α_{CI} or α_{AVCS}) has a dependence of $\frac{1}{m}$, where m is the number of hypotheses (i.e., the number of confidence intervals or sequences that can err). However, when we employ a filter M, the number of boosted e-values we may incorrectly evaluate becomes $|M| \leq m$. Therefore, we can replace m with $|M|$ in α_{CI} and α_{AVCS} while preserving the FDR control with additive Monte-Carlo error in Propositions 1 and 2. The effect on the confidence mechanism is that it becomes less conservative, which leads to higher power.

We close this section by formalizing the effect of filtering on the FDR control guarantee of the e-BH-CC procedure, whose proof can be found in Appendix A.4.

Proposition 3. *Let $\mathcal{R}(\boldsymbol{e})$ be the original e-BH rejection set and $\mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$ be the boosted e-BH rejection set, both at level $\alpha \in (0, 1)$. Then for all sets $\mathcal{S}$ such that $\mathcal{R}(\boldsymbol{e}) \subseteq \mathcal{S} \subseteq \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$, $\text{FDR}[\mathcal{S}] \leq \alpha$.*

3.3 Connection to dBH

As briefly discussed earlier, Fithian and Lei (2022) propose a method named dBH, which uses conditional calibration to adjust the p-value threshold in BH, so as to control the FDR under dependence. Concretely, supposing access to a sufficient statistic S_j for each $j \in [m]$ such that one can sample from $\boldsymbol{p} := (p_1, \ldots, p_m) \,|\, S_j$ under the null hypothesis, it finds the critical value $\widehat{\tau}_j$ (via numerical integration):[3]

[3] This is a simplified version of the dBH procedure. In the original formulation, the threshold is written as a function of the tuning parameter c.

$$\widehat{\tau}_j = \sup\left\{c \in (0,1) : \mathbb{E}_{H_j}\left[\frac{\mathbb{1}\{p_j \le \widehat{\tau}_j\}}{\widehat{r}_j(\boldsymbol{p})} \,\middle|\, S_j\right] \le \frac{\alpha}{m}\right\}, \tag{10}$$

where $\widehat{r}_j(\boldsymbol{p})$ is an estimate of $|\{j\} \cup \{k \in [m] : p_k \le \widehat{\tau}_k\}|$ (not to be confused with $\widehat{\mathcal{R}}_j(\boldsymbol{e})$). The adjusted p-value threshold then yields a selection set $\widehat{\mathcal{R}}^+ = \{j \in [m] : p_j \le \widehat{\tau}_j\}$. Note that the size of $\mathcal{R}^+$ may differ from the estimates $\widehat{r}_j(\boldsymbol{p})$, so the construction of $\widehat{\tau}_j$ does not necessarily imply FDR control. The dBH procedure takes additional pruning steps to ensure FDR control:

(1) if $|\widehat{\mathcal{R}}^+| \ge \widehat{r}_j(\boldsymbol{p})$ for all $j \in [m]$, then stop the procedure and return $\widehat{\mathcal{R}}^+$;

(2) if there exists $j \in [m]$ such that $|\widehat{\mathcal{R}}^+| < \widehat{r}_j(\boldsymbol{p})$, then generate $U_1, \dots, U_m \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$ that is independent of everything else, and return

$$\mathcal{R}^{\text{dBH}} := \left\{j \in \widehat{\mathcal{R}}^+ : U_j \le \tfrac{r^*}{\widehat{r}_j(\boldsymbol{p})}\right\}, \text{ where } r^* = \max\left\{r \in [m] : \left|\left\{j \in \widehat{\mathcal{R}}^+ : U_j \le \tfrac{r}{\widehat{r}_j(\boldsymbol{p})}\right\}\right| \ge r\right\}.$$

As we can see here, dBH makes use of conditional calibration to achieve the FDR control, while in our framework, we start with a FDR-controlling selection set and use conditional calibration to increase the power. Additionally, the dBH procedure involves additional pruning steps that introduce external randomness, while our method does not.

Before closing this section, we point out an interesting fact about dBH—it actually has an e-BH interpretation, based on which the external randomness can be avoided (with potential loss of power). To see this, we write

$$e_j^{\text{dBH}} = \frac{m\mathbb{1}\{p_j \le \widehat{\tau}_j\}}{\alpha \widehat{R}_j(\boldsymbol{p})}. \tag{11}$$

For simplicity, we assume that the conditional expectation in (10) is less than or equal to α/m at the critical value (otherwise we can replace the "$\le$" with "$<$" as in our e-BH-CC procedure). As a consequence, $\mathbb{E}[e_j^{\text{dBH}}] \le 1$ for any $j \in \mathcal{H}_0$, which means that e_j^{dBH} is a valid e-value. With this observation, we can apply the e-BH procedure to the e-values and obtain a deterministic selection set with FDR control. One can also check that applying e-BH to the "dBH e-values" in (11) is equivalent to dBH with all the U_j's replaced by 1; moreover, the dBH proceudure is equivalent to applying the e-BH procedure to $(e_1^{\text{dBH}}/U_1, \dots, e_m^{\text{dBH}}/U_m)$—this is the pe-BH procedure in Ignatiadis et al. (2023) or equivalently the U-eBH procedure in Xu and Ramdas (2023a). Such a connection has also been noticed under a different context in Jin and Candès (2023). We formalize the equivalence in the following proposition, with its proof deferred to Appendix A.5.

Proposition 4. *The e-BH procedure applied to $(e_1^{\text{dBH}}/U_1, \dots, e_m^{\text{dBH}}/U_m)$ is equivalent to the dBH procedure, where e_j^{dBH}'s are as defined in (11). In particular, if we replace all the U_j's by 1, then dBH is equivalent to e-BH applied to $(e_1^{\text{dBH}}, \dots, e_m^{\text{dBH}})$.*

4 Example: parametric testing

First, we study the relatively simple task of identifying which components of a Gaussian random variable have nonzero mean. Given an m-dimensional z-statistic $Z \sim \mathcal{N}_m(\mu, \Sigma)$ with $\Sigma \succ 0$, we define the null hypotheses $H_j : \mu_j = 0$ and their corresponding one-sided alternative $H_j^{\text{alt}} : \mu_j > 0$. Fithian and Lei (2022) consider this problem for both known and unknown covariance matrix Σ as an example application of dBH, their conditionally-calibrated correction to the BH procedure. After constructing the relevant e-values for testing against $H_1, \dots, H_m$, we closely follow their work in Sections 3.1 and 3.2 as their conditioning statistic plays the role of our sufficient statistic for the e-value resampling step.

4.1 Multivariate z-statistics

Consider $Z \sim \mathcal{N}_m(\mu, \Sigma)$ with Σ known. Without loss of generality, we assume $\Sigma_{jj} = 1$ for all j. Vovk and Wang (2021) observe that for $a_j \ne 0$,

$$e_j = \exp(a_j Z_j - a_j^2/2) \tag{12}$$

is a valid e-value with respect to H_j. e_j is the likelihood ratio test (LRT) statistic for H_j versus the point alternative $H_j^{(a_j)}\colon \mu_j = a_j$. The choice of the hyperparameter a_j can be either fixed *a priori* or derived from an independent hold-out set and will not affect the validity of the e-value. However, different values of a_j will lead to varying levels of power. The intuition from the Neyman-Pearson lemma is that the choice of a_j that matches with the true data-generating distribution under the alternative. When the alternative hypothesis for H_j is composite, our choice of a_j may lead to e_j being far from optimally powerful—but it is impossible to be certain without assuming we know the alternative distribution. In these cases, the practitioner may attempt to learn the best a_j independently from a separate, independent dataset. We run simulations, detailed in Section 4.3, where our method is applied for e-values constructed using correctly specified "exact" LRT e-values as well as a range of misspecified LRT e-values. We find that when estimating μ using a hold-out set, the performance is quite similar to the correctly specified e-values, so we do not present those results.

In order to use conditional calibration, we define the sufficient statistic $S_j = Z_{-j} - \Sigma_{-j,j}Z_j$ for each $j \in [m]$, as in Fithian and Lei (2022). To resample $\widetilde{e}_1, \ldots, \widetilde{e}_k \mid S_j$, it suffices to resample from the conditional joint distribution $Z \mid S_j$ (under H_j) and compute (12) for each j. We claim that we can sample $Y \sim \mathcal{N}(0, \Sigma_{jj})$ and construct the z-statistic $\widetilde{Z}^{(j)} = (\widetilde{Z}_1^{(j)}, \ldots, \widetilde{Z}_m^{(j)})$ such that

$$\widetilde{Z}_j^{(j)} = Y, \quad \widetilde{Z}_{-j}^{(j)} = S_j + \Sigma_{-j,j}Y \tag{13}$$

in order to obtain a resample from $Z \mid S_j$. The following proposition formalizes the claim.

Proposition 5. *For each $j \in [m]$, choose sufficient statistic $S_j = Z_{-j} - \Sigma_{-j,j}Z_j$ and resample the m-dimensional z-statistic $\widetilde{Z}^{(j)}$ from $Z \mid S_j$ as written in (13). Define the resampled e-values $\widetilde{e}_k^{(j)} = \exp(a_k\widetilde{Z}_k^{(j)} - a_k^2/2)$ for each $k \in [m]$. Then*

$$\left(\widetilde{e}_1^{(j)}, \ldots, \widetilde{e}_m^{(j)}\right) \mid S_j \sim (e_1, \ldots, e_m) \mid S_j$$

under H_j.

Furthermore, independent samples of $Y \sim \mathcal{N}(0, \Sigma_{jj})$ lead to independent resamples from $(e_1, \ldots, e_m) \mid S_j$.

Proof. It is sufficient to show that the resample $\widetilde{Z}$ follows the conditional joint distribution $Z \mid S_j$ under $H_j\colon \mu_j = 0$. Since $\mathrm{Cov}(Z_j, S_j) = 0$, the two entities are indepdendent. Thus, the independently sampled $\widetilde{Z}_j = Y \sim \mathcal{N}(0, \Sigma_{jj})$ follows $Z_j \mid S_j$ automatically.

Define the deterministic function $g\colon \mathbb{R} \times \mathbb{R}^{m-1} \to \mathbb{R}^{m-1}$ such that $g(z, S) = S + \Sigma_{-j,j}z$. Then $Z_{-j} = g(Z_j, S_j)$ and $\widetilde{Z}_{-j} = g(Y, S_j)$ (by (13)). Therefore,

$$Z_j \mid S_j \sim Y \mid S_j \implies (Z_{-j}, Z_j) \mid S_j \sim (\widetilde{Z}_{-j}, \widetilde{Z}_j) \mid S_j$$

and we conclude as desired. □

4.2 Multivariate t-statistics

When the covariance matrix Σ is instead partially or totally unknown, the t-statistic replaces its z-statistic counterpart. Assume that $Z \sim \mathcal{N}_m(\mu, \Sigma)$, where $\Sigma = \sigma^2\Psi$ with $\Psi \succ 0$ known but $\sigma \in \mathbb{R}^+$ unknown. Furthermore, assume we have access to an auxiliary independent vector $W \sim \mathcal{N}_{n-m}(0, \sigma^2 I_{n-m})$ for estimating σ^2. Then, under the same null hypotheses $H_1, \ldots, H_m$, the t-statistics used for testing is

$$T_j := \frac{Z_j}{\sqrt{\widehat{\sigma}^2\Psi_{jj}}} \sim t_{n-m},$$

where $(n-m)\widehat{\sigma}^2 = \|W\|^2 \sim \sigma^2\chi^2_{n-m}$. Here, t_{n-m} and χ^2_{n-m} refers to the student-t and χ^2 distributions with $n-m$ degress of freedom.

For e-values, we can again construct a likelihood ratio test statistic per j. Under the point alternative $H_j^{(a_j)}\colon \mu_j = a_j$, $T_j \sim t_{n-m}(a_j)$, the noncentral t-distribution with noncentrality parameter a_j and $n-m$ degrees of freedom; this distribution is a generalization of the regular t-distribution under the relation $t_{n-m}(0) \overset{\mathrm{d}}{=} t_{n-m}$. Denoting the density of $t_{n-m}(a)$ as $f_{t_{n-m},a}$, we then define for each H_j the LRT e-value

$$e_j = \frac{f_{t_{n-m},a_j}(T_j)}{f_{t_{n-m},0}(T_j)}, \tag{14}$$

where again the choice of a_j can be chosen *a priori* or through sample splitting.

For the conditioning statistic S_j, Fithian and Lei (2022) construct the pair (U_j, V_j) with the components defined as follows:

$$\begin{aligned} U_j &= Z_{-j} - \Psi_{-j,j}\Psi_{jj}^{-1} Z_j, \\ V_j &= \|W\|^2 + \frac{Z_j^2}{\Psi_{jj}}. \end{aligned} \tag{15}$$

By choosing this statistic, we again have $S_j \perp\!\!\!\perp T_j$; in fact, U_j, V_j, and T_j are all mutually independent. In addition, it can be shown that T_{-j} is a deterministic function of (T_j, U_j, V_j):

$$T_k = U_{jk}\sqrt{\frac{n-m+T_j^2}{\Psi_{kk}V_j}} + \frac{\Psi_{kj}}{\Psi_{jj}}T_j \quad \text{for } k \neq j. \tag{16}$$

As a direct consequence of (16), we can resample from $T \mid S_j$ by sampling $Y \sim t_{n-m}$ independent from S_j, setting $\widetilde{T}_j = Y$, and setting $\widetilde{T}_{-j}$ as the function of $\widetilde{T}_j, S_j$ above. The LRT e-values can be calculated per-component. We formalize the correctness of this resampling scheme in the following proposition.

Proposition 6. *For each $j \in [m]$, choose sufficient statistic $S_j = (U_j, V_j)$ as defined in (15). Using U_j, V_j, and $\widetilde{T}_j = Y \sim t_{n-m}$ (independent from S_j), construct $\widetilde{T}_{-j}$ as per (16). Then $\widetilde{T} \mid S_j \sim T \mid S_j$ and*

$$\left(\widetilde{e}_1^{(j)}, \ldots, \widetilde{e}_m^{(j)}\right) \mid S_j \sim (e_1, \ldots, e_m) \mid S_j$$

under H_j, where e_k and $\widetilde{e}_k^{(j)}$ is constructed as the component-wise LRT by using (14) on T and $\widetilde{T}$, respectively.

Furthermore, independent samples of $Y \sim t_{n-m}$ lead to independent resamples from $(e_1, \ldots, e_m) \mid S_j$.

Proof. We can proceed similarly to the proof of Proposition 5. As Fithian and Lei (2022) show, $S_j = (U_j, V_j)$ is independent to T_j. Therefore, the independently sampled $\widetilde{T}_j = Y \sim t_{n-m}$ is distributed as $T_j \mid S_j$. Noting that $\widetilde{T}_{-j}$ depends on $\widetilde{T}_j, S_j$ in the same exact deterministic way as T_{-j} depends on T_j, S_j, we conclude that

$$(T_{-j}, T_j) \mid S_j \sim (\widetilde{T}_{-j}, \widetilde{T}_j) \mid S_j.$$

The proposition is then immediate. □

4.3 Simulation studies

We show the power improvement given by e-BH-CC on the LRT e-values for both the z-testing and t-testing problems.

Settings. We take $m = 100$ and the set of nonnull $\mathcal{H}_1 = \{1, \ldots, 10\}$, so that

$$\mu = (\underbrace{A, A, \ldots, A}_{10}, 0, \ldots, 0) \in \mathbb{R}^{100},$$

where A is a constant determining the signal strength. The covariance matrix has the form $\Sigma_{ij} = \rho^{|i-j|}$ for any $i, j \in [m]$. The Gaussian data is thus generated from $\mathcal{N}(\mu, \Sigma)$. Depending on the specific testing problem, we will choose various ranges for the signal strength A and covariance parameter ρ.

Methods. We compare e-BH and e-BH-CC, both at level $\alpha = 0.05$. As a reference, we also run BH at α on the one-sided p-values derived from the z-statistics and t-statistics. Note that BH does not have FDR control guarantees when $\rho < 0$, which we consider in the t-testing simulations, so it is not truly comparable with our e-value methods (which do guarantee FDR control).

Computational details. We use a hybrid AVCS formulation to control the Monte-Carlo error. This is a computational trick which balances the conservativeness of the exact finite-sample AVCS with the asymptotic AVCS (although we lose any meaningful statement regarding the limiting FDR). The first 3000 samples, batched in sizes of 100 and scaled to the unit interval, are used to form the hedged capital

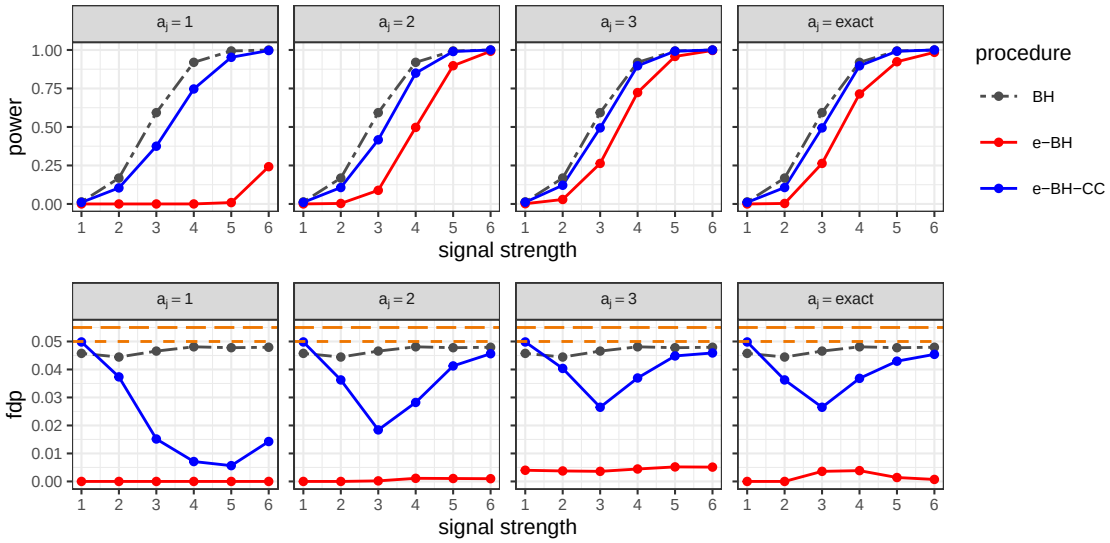

Figure 2: Realized power and FDP of the simulated experiments for z-testing. Each plot contains the averaged metrics over 1000 replications. The short and long orange dashed lines in the FDP plots represent the target FDR (0.05) and the FDR bound with Monte-Carlo error (0.055), respectively.

confidence sequence (HCCS) as proposed in Waudby-Smith and Ramdas (2024). After this point, the next samples, also batched in sizes of 100, are used to construct the asymptotic AVCS described in Theorem 2.2, Waudby-Smith et al. (2021). If after 5000 samples the resulting AVCS still contains zero, we stop early and fail to boost the e-value.

In addition, we employ a filter to cut down the number of e-values which we attempt to boost. For all experiments, we filter out all j such that $e_j = 0$, since they see no benefit from boosting. Furthermore, we use the filter

$$M = \{j : p_j \leq 3\alpha\},$$

where p_j are the one-sided p-values formed from the z-statistics or t-statistics.

We chooose $\alpha_0 = 0.1 \cdot \alpha = 0.005$ and use α_{AVCS} as described in Algorithm 3, replacing m with the size of the filter. This is in line with our target FDR (accounting for Monte-Carlo error) of 0.05. Even though the asymptotic AVCS only bounds above the limit supremum of the FDR at $\alpha + \alpha_0 = 0.055$, we see that the FDR is empirically controlled at 0.05.

4.3.1 Experiments for z-testing

We vary $A \in \{1, 2, \ldots, 6\}$ and set $\rho = 0.5$. Four different constructions of LRT e-values are considered: the first three constructions let $a_j = a$, $\forall j \in [m]$ for $a \in \{1, 2, 3\}$; for the last construction, we use a correctly specified a_j (setting it equal to A). The mean power and FDP curves, taken over 1000 replications, are shown in Figure 2.

For all choices of a_j, we see a major improvement in power comparing e-BH-CC to base e-BH. The power improvement is uniform over all possible signal strengths, but is especially large when a_j is incorrectly specified. For example, when $a_j = 1$, we see that e-BH has minimal power even in the large signal experiments. e-BH-CC immensely outperforms e-BH in this situation.

It is worth noting that in these experiments, BH still controls the FDR as the p-values are PRDS (Benjamini and Yekutieli, 2001). Thus, we can intrepret the power gap between BH and e-BH as the power loss from translating the multiple testing problem from p-values to e-values. The e-BH-CC power curve then demonstrates how much of the power loss can be reclaimed. In all cases, we see that the power of e-BH-CC is quite comparable to that of BH. The power reclamation is even more evident in the misspecified case $a_j = 1$, as mentioned previously. One takeaway is that when a_j is chosen poorly through fixed means or estimation, e-BH-CC can still perform powerfully, giving an added margin of safety regarding parameter misspecification.

Lastly, it is clear that the FDR is controlled empirically at $\alpha + \alpha_0$ (even at α) for all settings. The

realized FDP of e-BH-CC is generally much higher than that of base e-BH (which is close to zero), affirming that we are able to use more of the FDR budget by boosting e-BH.

4.3.2 Experiments for t-testing

We vary $\rho \in \{0.9, -0.9\}$ and degrees of freedom $n-m \in \{5, 50\}$ to visualize the effect of e-BH-CC in both heavy and light tailed settings. For $n-m=5$ (the heavy tailed setting), we vary $A \in \{2, 2.5, 3, \ldots, 6.5\}$, while for $n-m=50$ (the light tailed setting), we vary $A \in \{1, 1.5, 2, \ldots, 5.5\}$. For each choice of hyperparameters, we assume that a_j is chosen correctly, setting it equal to the signal strength A. The mean power and FDP curves, taken over 1000 replications, are shown in Figure 3.

Again, we see that in all settings e-BH-CC improves upon e-BH. The improvement is notable when the signals are stronger, although when the signals are weak all three methods (including BH) suffer in power. Note that in these settings, BH does not necessarily control the FDR; however, we still find it pertinent to consider the gap between it and e-BH a loss of power. In terms of reclaiming the power gap between BH and e-BH, e-BH-CC does quite well and is even comparable to BH in the light-tailed experiments. To conclude, we again note that e-BH-CC controls the FDR empirically and shows a better usage of the FDR budget compared to e-BH.

5 Example: knockoffs

5.1 Conditional independence testing

We turn to the problem of *feature selection*, where the goal is to discover which of the covariates $X = (X_1, X_2, \ldots, X_m)$ are significant to the value of the outcome variable Y. We can encode the notion of X_j being significant to the outcome with the following *conditional independence* hypothesis:

$$H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}.$$

These can serve as the null hypotheses: if H_j is true, then the value of X_j has no effect on the outcome, controlling for all other variables. We call X_j a *null* variable when H_j is true and *nonnull* when H_j is false. The null hypothesis set $\mathcal{H}_0$ corresponds to the indices of the null covariates, and analogously $\mathcal{H}_1$ corresponds to the nonnull covariates. Our goal is to reject a subset of $[m]$ in a way which controls the FDR, i.e., there is little intersection with $\mathcal{H}_0$.

One popular approach for testing conditional independence is to assume the *model-X framework*. In this framework, the covariate-outcome tuple $(X_1, \ldots, X_m, Y)$ is interpreted as a draw from some joint

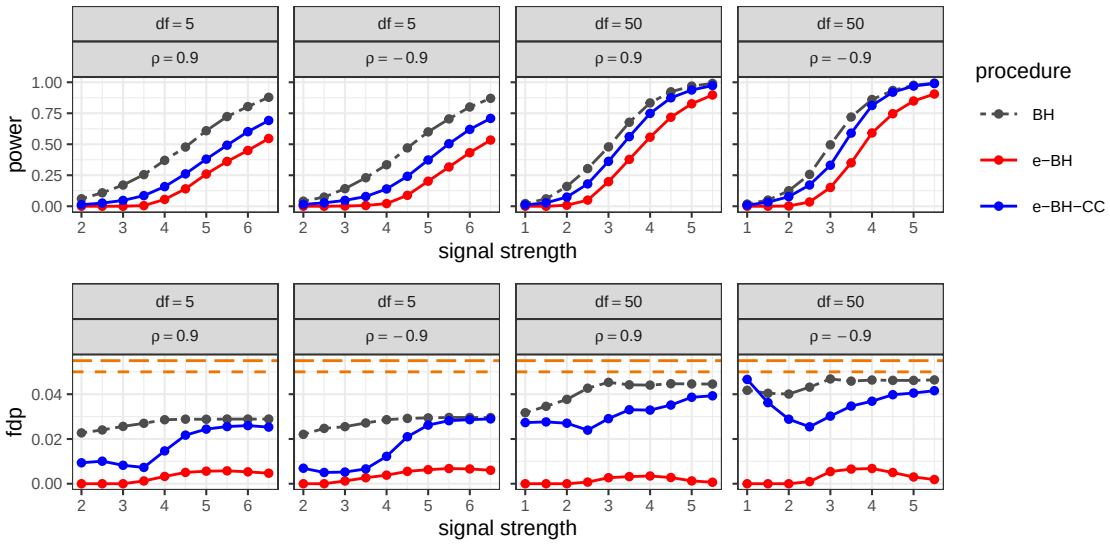

Figure 3: Realized power and FDP of the simulated experiments for t-testing. The details are otherwise the same as in Figure 2.

distribution $\mathbb{P}_{XY} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$, where $\mathbb{P}_X$ (the covariate distribution) is assumed to be known (or reasonably well-approximated) while the model $\mathbb{P}_{Y|X}$ is totally unknown. Under these assumptions, one can use a multiple testing procedure known as the *model-X knockoff filter* (Candès et al., 2018) to conduct feature selection with a provable exact-sample FDR control guarantee. We give a brief exposition of the procedure in the following section with the goal of showing its relevance to e-values and e-BH-CC.

5.2 Model-X knockoffs

Given the design matrix $\boldsymbol{X} = (X_{i1}, \dots, X_{im})_{i\in[n]} \in \mathbb{R}^{n\times m}$ and outcomes $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ in the dataset, the model-X knockoff filter uses the practitioner's knowledge of the covariate distribution $\mathbb{P}_X$ to construct "knockoff variables" $\widetilde{\boldsymbol{X}} = (\widetilde{X}_{i1}, \dots, \widetilde{X}_{im})_{i\in[n]}$. These knockoff variables must satisfy an independence condition $\boldsymbol{Y} \perp\!\!\!\perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{X}$ as well as the "swap" condition:

$$(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}) \sim (\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}) \;\; \forall j \in [m],$$

that is, the joint distribution of $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ is preserved after swapping the positions of the columns $\boldsymbol{X}_j$ and $\widetilde{\boldsymbol{X}}_j$. Hence, $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are quite similar in terms of their dependency structure, but the latter has no significance to the outcome that is not already expressed through $\boldsymbol{X}$. Hence, conditional on $\boldsymbol{X}$, $\widetilde{\boldsymbol{X}}$ are knockoff covariates which carry no value (in predicting $\boldsymbol{Y}$).

At a high level, access to these knockoff variables provides a way to appropriately calibrate any feature importance measure of X_j. The procedure uses the augmented design matrix $[\boldsymbol{X}, \widetilde{\boldsymbol{X}}]$ and the outcome $\boldsymbol{Y}$ to construct a feature importance vector $W = \mathcal{W}([\boldsymbol{X}, \widetilde{\boldsymbol{X}}], \boldsymbol{Y}) \in \mathbb{R}^m$ (where W_j corresponds to X_j and its knockoff) using some algorithm $\mathcal{W}(\cdot)$ with the property that switching $\boldsymbol{X}_j$ and $\widetilde{\boldsymbol{X}}_j$ in the augmented design matrix flips the sign of the resulting W_j. In general, such statistics are computed by finding individual feature importances for X_j and $\widetilde{X}_j$ and taking their difference. As noted in Candès et al. (2018), the signs of $\{W_j : j \in \mathcal{H}_0\}$ conditioned on their magnitudes $\{|W_j| : j \in \mathcal{H}_0\}$ are i.i.d. coin flips. In contrast, W_j for nonnull X_j would tend to be positive (signifying more importance of X_j than $\widetilde{X}_j$). This aligns with the intuition of $\widetilde{X}_j$ as "negative controls": when X_j and $\widetilde{X}_j$ are both insignificant, the feature importance statistic is symmetric around 0, but when X_j is significant W_j will be positively skewed. Using the feature importances W, we select the set of features:

$$\mathcal{R}^{\text{kn}} = \{j : W_j \geq T\}, \qquad \text{where } T := \inf\left\{t > 0 : \frac{1 + \sum_{j\in[m]} \mathbb{1}\{W_j \leq -t\}}{\sum_{k\in[m]} \mathbb{1}\{W_k \geq t\}} \leq \alpha\right\}. \tag{17}$$

The threshold T is defined to be the smallest $t > 0$ such that an estimate of the FDR (which is constructed using our intuition on the null i.i.d. coin flips above) is controlled by α. For a rigorous proof of FDR control, see Candès et al. (2018); Barber and Candès (2015).

Although the knockoff filter is an elegant and effective approach to feature selection, it still has a few weaknesses. One such weakness, known as the *threshold phenomenon*, occurs when the number of features with non-negligiable significance is less than $1/\alpha$. In this sparse setting, the knockoffs procedure experiences a drastic loss in power. An explanation for this phenomenon is in the definition of T, where we compare an estimator of the FDP (false discovery proportion) against α. The denominator measures the size of the rejection set when using t, while the numerator is at least 1; when the denominator is less than $1/\alpha$, $\mathcal{R}^{\text{kn}}$ is forced to be empty. In practice, one still sees the procedure suffer from powerlessness when the proportion of significant features is generally sparse, even when $|\mathcal{H}_1|$ is technically above $1/\alpha$.

5.2.1 Derandomizing knockoffs with e-values

In addition to the threshold phenomenon, another weakness of the knockoff filter is its high selection variability. Sampling the knockoff matrix $\widetilde{\boldsymbol{X}}$ introduces extraneous randomness into the procedure, so two different runs of the knockoff filter (with two distinct knockoff matrices) may result in wildly different rejection sets. For practical purposes, this is a major detraction—for example, two scientists studying the same feature selection problem with the same dataset may conclude totally different sets of significant covariates. Hence, for scientific reproducibility, a way to derandomize the knockoff filter was highly sought-after.

Ren and Barber (2024) solve the variability issue by proposing a derandomized version of the knockoff filter. They begin by connecting the knockoff filter with the e-BH procedure through defining

$$e_j := m \cdot \frac{\mathbb{1}\{W_j \geq T\}}{1 + \sum_{k \in [m]} \mathbb{1}\{W_j \leq -T\}}, \tag{18}$$

where W_j is the feature importance for X_j and T is the rejection threshold in (17). It can be shown that $\sum_{j \in \mathcal{H}_0} \mathbb{E}[e_j] = m$, implying that $e_1, \ldots, e_m$ are generalized e-values (Definition 1). Running e-BH on these e-values will return a rejection set with FDR control, and the authors show that this rejection set is identical to the output of the knockoff filter using the same knockoff matrix.

The procedure to derandomize the knockoff filter is as follows. After choosing a hyperparameter $\alpha_{\text{kn}} \in (0, 1)$, sample d knockoff samples $(\widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(d)})$. For each knockoff $\widetilde{\boldsymbol{X}}^{(k)}$, we run the knockoff filter with the feature importances $W^{(k)} = \mathcal{W}([\boldsymbol{X}, \widetilde{\boldsymbol{X}}^{(k)}], \boldsymbol{Y})$ at level α_{kn}, which results in the threshold

$$T^{(k)} := \inf \left\{ t > 0 \colon \frac{1 + \sum_{j \in [m]} \mathbb{1}\left\{W_j^{(k)} \leq -t\right\}}{\sum_{k \in [m]} \mathbb{1}\left\{W_k^{(k)} \geq t\right\}} \leq \alpha_{\text{kn}} \right\}.$$

We then use (18) to construct the generalized e-values for the kth run of the knockoff filter:

$$e_j^{(k)} = m \cdot \frac{\mathbb{1}\left\{W_j^{(k)} \geq T^{(k)}\right\}}{1 + \sum_{k \in [m]} \mathbb{1}\left\{W_j^{(k)} \leq -T^{(k)}\right\}}, \quad \forall j \in [m]. \tag{19}$$

For each feature j, we can construct a derandomized e-value by averaging $e_j^{(k)}$ over all d runs of the knockoff filter. The averaged e-value, formally defined as

$$\bar{e}_j := \frac{1}{d} \sum_{k \in [d]} e_j^{(k)}, \tag{20}$$

exhibits more stability as the extraneous randomness from each run of knockoffs is averaged out. Since taking the average preserves (generalized) e-value validity, we can run e-BH on $(\bar{e}_1, \ldots, \bar{e}_m)$ to get a rejection set while controlling the FDR at level α.

The hyperparameter α_{kn} does not necessarily have to equal α (the desired FDR control), and Ren and Barber (2024) suggest having the former depend on the latter through the choice $\alpha_{\text{kn}} = \alpha/2$ in order to achieve good power. Regardless of the choice of α_{kn}, the resulting $\bar{e}_1, \ldots, \bar{e}_m$ will be valid e-values. In Figures 4, 5, and 6, we show the effect of conditional calibration on power and FDR for multiple choices of α_{kn}. In addition, there are other slight improvements which can affect the power of derandomized knockoffs—we discuss our implementation for the simulations in more detail in Section 5.4.

Unfortunately, even with these added techniques, the power of derandomized knockoffs procedure tends to suffer relative to that of the original method. This is due to the fact that the knockoff e-values in (18) (which is equivalent to the original method) are "tight" in the sense that the inequality (i) in (1) is nearly tight. However, the average of tight e-values is no longer tight, leading to a loss of power from e-BH. This power drop is most apparent in regimes with low-to-moderate signal strength; one can intrepret this as looseness in the FDR control inequality translating to signals becoming relatively weaker. To add to this, the gap becomes noticeably larger the more derandomization runs d are used.

Therefore, we can identify two settings for conditional independence testing where the knockoff filter, as an e-value procedure, suffers from power loss: sparse signals and derandomization. However, we can reclaim the power loss and fill in the gap by using conditional calibration to boost the derandomized knockoffs procedure—even in settings where the threshold phenomenon may happen. In the following subsection, we will describe how to implement e-BH-CC for derandomized model-X knockoffs.

5.3 Conditional calibration for derandomized knockoffs

The choice of conditioning statistic S_j must allow the i.i.d. resampling of $(e_1, \ldots, e_m)$ under the null conditional independence hypothesis H_j. Intuitively, it is sufficient if we can resample the design matrix itself in a way that fulfills H_j; i.e., the jth column of the resampled matrix is no longer significant to

the outcome (conditional on the other columns). By taking this resampled design matrix as the true $\boldsymbol{X}$, we can sample its knockoff matrix d times to construct the derandomized e-values. The jth e-value will correspond to a true null H_j by the nature of the resampled design matrix.

Resampling the design matrix is a crucial step in the *conditional randomization test* (CRT), which gives a p-value for testing against H_j in the model-X framework (Candès et al., 2018). In the first step of the CRT, the resampled matrix is constructed by drawing a new copy of the jth column from $\mathbb{P}_{X_j|X_{-j}}$, which is known through the model-X assumption; it is then concatenated with the non-jth columns of the original design matrix and combined with the outcome vector $\boldsymbol{Y}$ to simulate the original dataset. This formulation of the CRT leads us to choose $S_j = (\boldsymbol{X}_{-j}, \boldsymbol{Y})$ as the sufficient statistic.

Like in the CRT, we can construct a resample $\boldsymbol{X}'$ of the design matrix $\boldsymbol{X}$ conditional on S_j: we invoke our knowledge of $\mathbb{P}_{X_j|X_{-j}}$ to let $\boldsymbol{X}'_j$ be samples from the conditional distribution and assign to $\boldsymbol{X}'_{-j}$ the corresponding column values of $\boldsymbol{X}_{-j}$. We can use this resample as our starting point for the derandomized knockoffs procedure (with d derandomization runs), which outputs the averaged e-values $\widetilde{\bar{e}}_1, \dots, \widetilde{\bar{e}}_m$.

The following proposition states the correctness of the afore-described resampling procedure in producing i.i.d. resamples from $(\bar{e}_1, \dots, \bar{e}_m) \mid S_j$.

Proposition 7. *Fix $\alpha \in (0,1)$, $\alpha_{\mathrm{kn}} \in (0,1)$, and $d \in \mathbb{N}$. For each $j \in [m]$, choose sufficient statistic $S_j = (\boldsymbol{X}_{-j}, \boldsymbol{Y})$. As in the CRT (Candès et al., 2018), construct a resample $\boldsymbol{X}'$ of $\boldsymbol{X}$ by letting $\boldsymbol{X}'_{-j} = \boldsymbol{X}_{-j}$ and X'_{ij} be a sample from $\mathbb{P}_{X_{ij}|X_{i,-j}}$. Using the new, resampled dataset $(\boldsymbol{X}', \boldsymbol{Y})$, run the derandomized knockoffs procedure (with hyperparameters d and α_{kn}) at level α to get the e-values $\widetilde{\bar{e}}_1, \dots, \widetilde{\bar{e}}_m$, where $\widetilde{\bar{e}}_j$ is constructed as (19) and (20). Then*

$$(\widetilde{\bar{e}}_1, \dots, \widetilde{\bar{e}}_m) \mid S_j \sim (\bar{e}_1, \dots, \bar{e}_m) \mid S_j$$

under H_j. Furthermore, independent resamples $\boldsymbol{X}'$ will lead to independent $(\widetilde{\bar{e}}_1, \dots, \widetilde{\bar{e}}_m)$ conditional on S_j (also under H_j).

Proof. The proposition follows once we show the distributional equality between the old and resampled columns $\boldsymbol{X}_j$ and $\boldsymbol{X}'_j$ (conditional on $\boldsymbol{X}_{-j}$ and $\boldsymbol{Y}$) under the null H_j. The null states that $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}$. Since the resample was thus created without any extraneous knowledge of $\boldsymbol{Y}$ (outside of what is contained when conditioning on $\boldsymbol{X}_{-j}$), the desired property simplifies to

$$\boldsymbol{X}_j \mid \boldsymbol{X}_{-j} \sim \boldsymbol{X}'_j \mid \boldsymbol{X}_{-j}.$$

However, the above is evident since we resampled $\boldsymbol{X}'_j$ by the law $\otimes_{i=1}^n \mathbb{P}_{X_{ij}|X_{i,-j}}$. □

5.4 Simulation studies

We illustrate the power improvement obtained from using e-BH-CC on the derandomized knockoffs e-values through numerical experiments.

Settings. We generate the design matrix and outcome vector under the Gaussian linear model. Let $n = 500$ be the number of observations and $m = 200$ be the number of covariates (i.e., number of hypotheses). For each observation, the row of covariates are jointly drawn from $X \sim \mathcal{N}_m(\mu, \Sigma)$. In our experiments, we set $\mu = 0$ and $\Sigma_{ij} = 0.5^{|i-j|}$, for any $i, j \in [m]$.

We then conditionally draw the response by the Gaussian linear model $Y \mid X \sim \mathcal{N}_n(X^\top \beta, 1)$, giving one independent draw of the covariate-outcome pair (X, Y). We repeat this n times to obtain our dataset $(\boldsymbol{X}, \boldsymbol{Y})$. The coefficient vector β is formatted as

$$\beta = \left(\underbrace{0, \dots, 0}_{z}, \frac{A}{\sqrt{n}}, \underbrace{0, \dots, 0}_{z}, -\frac{A}{\sqrt{n}}, \dots \right) \in \mathbb{R}^m;$$

that is, every z zeroes is followed by one nonzero amplitude with alternating sign. Note that this directly translates to $|\mathcal{H}_1| = \lfloor \frac{m}{z+1} \rfloor$.

We will consider two separate regimes: the dense regime (with $z = 7$ and $|\mathcal{H}_1| = 25$ nonnull variables) and the sparse regime (with $z \in \{19, 20\}$ and $|\mathcal{H}_1| \in \{10, 9\}$). By running simulations for both regimes, we can demonstrate the power improvement from e-BH-CC even in the presence of the threshold phenomenon, as the target level will be set at $\alpha = 0.1$. In the dense regime, we will simulate for each $A \in \{3, 4, 5, 6\}$; in the sparse regime, we will simualte for each $A \in \{4, 5, 6, 7\}$. By doing so, we can capture the range of power over a variety of signal magnitudes.

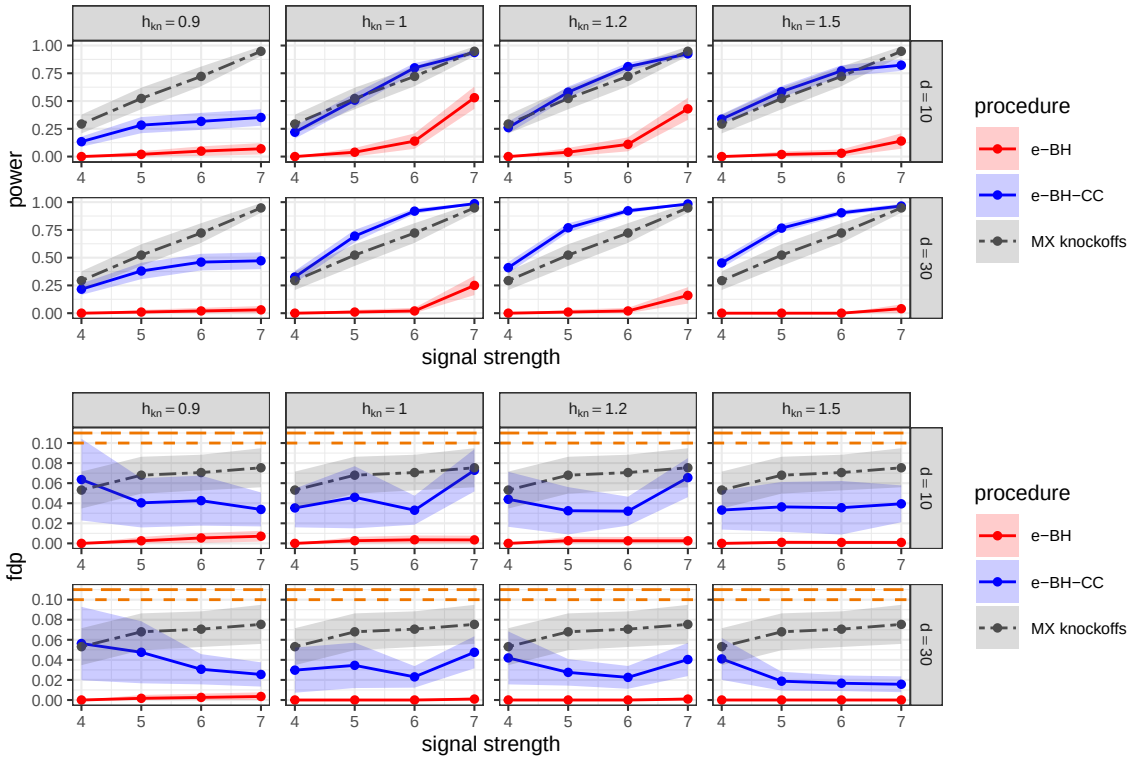

Figure 4: Realized power and FDP for the knockoffs simluations when β is sparse (specifically, $|\mathcal{H}_1| = 10$ and $\alpha = 0.1$, so this is *at* the threshold). Each facet contains the averaged metrics over 100 replications; shading indicates error bars. The short and long orange dashed lines in the FDP plots represent the target FDR (0.1) and the FDR bound with Monte-Carlo error (0.11), respectively.

Methods. The objective is to compare the derandomized knockoffs procedure before and after boosting at the target FDR level $\alpha = 0.1$, so we keep the hyperparameters and implementations the same between the two procedures. Thus, we find it prudent to only detail the basic derandomized knockoffs procedure. The Monte-Carlo details will only pertain to the boosted procedure.

For the dense β experiments, we use $d \in \{2, 10\}$ copies of the knockoff design matrix to derandomize. We run experiments for multiple values of the hyperparameter α_{kn} by choosing a factor $h_{\text{kn}} \in \{0.5, 0.75, 0.9, 1.0\}$ and defining $\alpha_{\text{kn}} = h_{\text{kn}}\alpha = 0.1h_{\text{kn}}$. In contrast, for the sparse β experiments, we use $d \in \{10, 30\}$ copies of the knockoff design matrix to derandomize and instead consider $h_{\text{kn}} \in \{0.9, 1.0, 1.2, 1.5\}$. We consider a range of h_{kn} to show that e-BH-CC empirically leads to uniform improvements in power. Further, to combat the threshold phenomenon it intuitively helps to have $\alpha_{\text{kn}} > \alpha$ so that the threshold T is lower and the number of nonzero e-values is consequently higher (to demonstrate this, we let h_{kn} range from below to above 1). The reason for $\alpha_{\text{kn}} \leq \alpha$ in the dense β case comes from the extensive experiments in the original derandomized knockoffs paper (Ren and Barber, 2024).

Knockoffs details. When sampling the knockoff design matrix, we use Equation (3.2) in Candès et al. (2018), where the S-matrix is constructed via the minimum variance-based reconstructibility criterion (see Definition 3.1 in Spector and Janson (2022)). We compute the lasso-coefficient difference (Equation (3.7), Candès et al. (2018)) as the feature importance measure. To avoid a computational slowdown from having to repeat cross-validation steps for each call of LASSO regression, we use a separate independently-sampled dataset ($n = 500$) as a "hold-out dataset" to pre-compute the λ regularization parameter. This pre-computation is repeated for every replication per experiment.

We also use a variant of the threshold T in (17) which Ren and Barber (2024) argue lead to a uniform

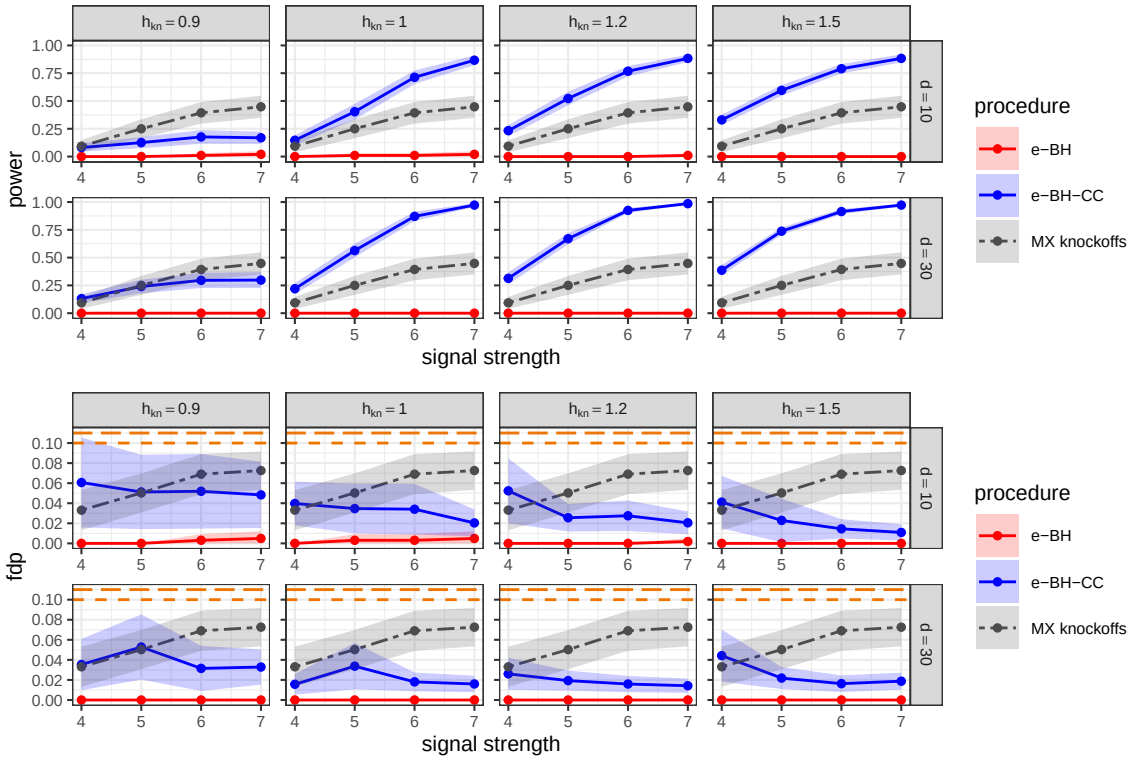

Figure 5: Realized power and FDP for the knockoffs simluations when β is sparse (specifically, $|\mathcal{H}_1| = 9$ and $\alpha = 0.1$, so this is *below* the threshold). The details are otherwise the same as in Figure 4.

improvement in power for derandomized knockoffs. The alternative threshold

$$T_{\text{alt}}^{(k)} := \inf\left\{t > 0 \colon \frac{1 + \sum_{j\in[m]} \mathbb{1}\left\{W_j^{(k)} \le -t\right\}}{\sum_{k\in[m]} \mathbb{1}\left\{W_k^{(k)} \ge t\right\}} \le \alpha_{\text{kn}} \text{ or } \sum_{j\in[m]} \mathbb{1}\left\{W_j \ge t\right\} < \frac{1}{\alpha_{\text{kn}}}\right\}$$

is an "early-stopping" version—when the original threshold is infinite (and all the e-values will be zero), using the alternative stopping time can only increase the e-values. We use $T_{\text{alt}}^{(k)}$ to give the base procedure some more power, while still observing that there is a large power gap to recover.

In our implementation, we frequently make use of the `knockpy` Python package, which is a Python implementation of the knockoffs procedure (Spector and Janson, 2022). We build on their functionality to implement derandomized knockoffs.

Computational details. We again use a hybrid AVCS formulation to control the Monte-Carlo error (see Section 4.3 for details). The first 1200 samples (in batches of 100) are used to construct the exact AVCS, while the next 800 samples are used to construct the asymptotic AVCS. After 2000 total samples, if the resulting AVCS still contains zero we stop early and fail to boost.

In addition, we will filter to cut down the number of e-values which we attempt to boost. For all experiments, we filter out all j such that $e_j = 0$, since they see no benefit from boosting. For $d = 30$, we will additionally filter using

$$M = \{j \colon \widehat{\beta}_j^{\text{LASSO}} \neq 0\} \cup \{j \colon p_j^{\text{regression}} \le \alpha\},$$

where $\widehat{\beta}^{\text{LASSO}}$ is the estimated coefficient vector obtained from LASSO regression on $(\boldsymbol{X}, \boldsymbol{Y})$ and $p_j^{\text{regression}}$ is the p-value corresponding to the jth covariate, derived from running OLS with intercept. The filter M attempts to capture covariates which are significant to the outcome while avoiding being too strict. Lastly, we chooose $\alpha_0 = 0.1 \cdot \alpha = 0.01$ and use α_{AVCS} as described in Algorithm 3, replacing m with the size of the filter.

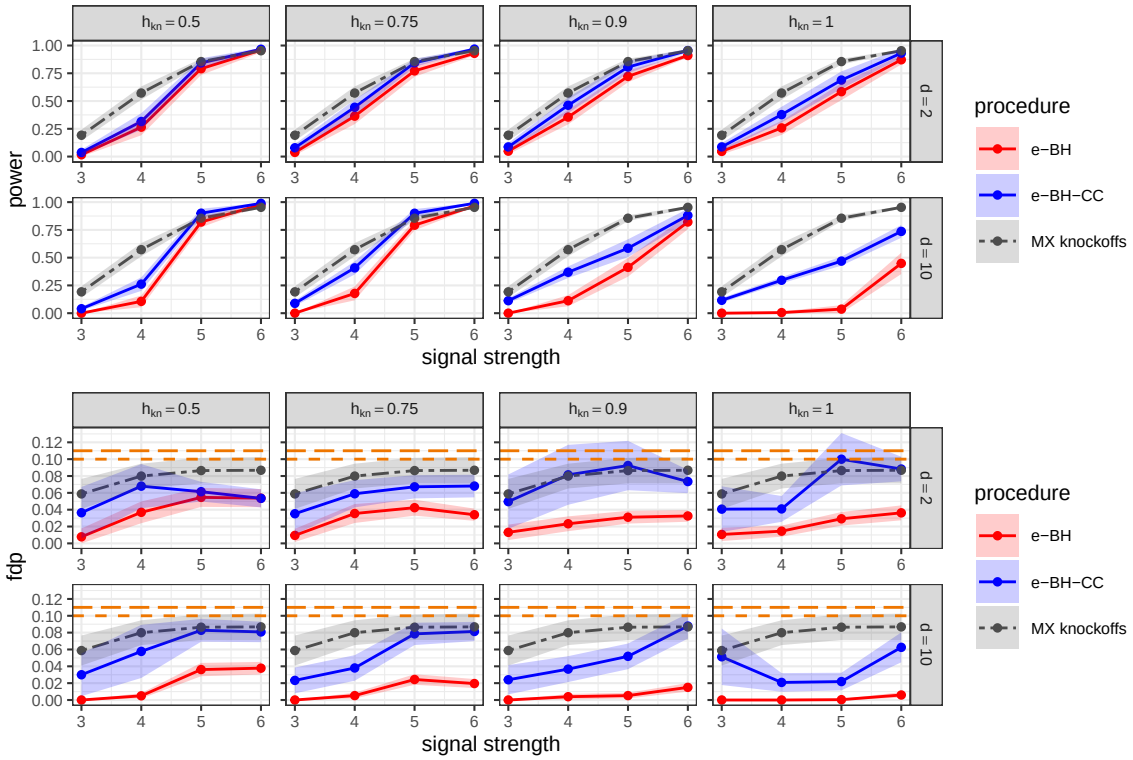

Figure 6: Realized power and FDP for the knockoffs simluations when β is dense ($|\mathcal{H}_1| = 25$). The details are otherwise the same as in Figure 4.

Experiments and results. We run 100 replications for each choice of A, z, d, h_{kn}. Each replication consists of constructing the basic derandomized knockoffs e-values and boosting these e-values using conditional calibration. In addition, for each choice of A and z, we run the original model-X knockoff filter (with the same implementation as derandomized knockoffs). This serves as a baseline and a demonstration of the power loss created by derandomization; the difference between the two power curves (shown in Figures 5, 4, and 6) of original knockoffs and derandomized knockoffs illustrates this power gap. We find that using e-BH-CC on the derandomized knockoffs e-values closes the gap, sometimes even exhibiting comparable-to-better power than the original knockoffs filter. However, we stress that the comparison between e-BH-CC and original knockoffs is technically an apples-to-oranges comparison, as the latter is a randomized procedure. For each procedure, we record its power and false discovery proportion (FDP) and plot their means per setting on the aforementioned figures.

The results of the dense setting experiment are shown in Figure 6. Even when $d = 2$, the drop in power exhibited by derandomized knockoffs is significant. However, e-BH-CC is able to improve the power *for all* choices of d and α_{kn}. The power drop, and subsequent power gain from boosting, are even more apparent when $d = 10$. The FDR is controlled well below the theoretical Monte-Carlo-adjusted error.

The results of the sparse setting experiments are shown in Figures 4 and 5. The setting of Figure 4, where $|\mathcal{H}_1| = 1/\alpha = 10$, experiences a major power drop over all signal amplitudes for both $d = 10, 30$. The resulting power gain from e-BH-CC is large—for $d = 10$, the power is comparable to the original knockoffs filter, while for $d = 30$, the power looks to be marginally better.

The setting of Figure 4, where the threshold phenomenon is in full swing (as $|\mathcal{H}_1| < 1/\alpha = 10$), shows an even larger contrast between the three procedures. When $h_{\text{kn}} \geq 1$ and for either $d \in \{10, 30\}$, derandomized knockoffs shows near 0 power. The original knockoffs procedure also exhibits lower power than in the case where $|\mathcal{H}_1| = 10$. Using e-BH-CC leads to an extremely large power gain over e-BH-CC, going from near-zero power to near-full power for high signal amplitudes. Furthermore, e-BH-CC has much better performance than the regular knockoffs filter. These results suggest that e-BH-CC does not experience a phase transition at the threshold in the same way that the knockoffs filter does. The power curves of e-BH-CC between Figures 4 and 5 differ slightly, while that of the knockoffs filter differ greatly, faltering when $|\mathcal{H}_1| < 1/\alpha$.

Finally, the results reiterate that even with the asymptotic AVCS trick, we still attain empirical FDR control at α uniformly over all choices of hyperparameters.

6 Example: conformalized outlier detection

Given a set of data, we consider the task of identifying the units whose distributions differ from the that of a reference dataset. This problem is known as outlier detection, and it is a fundamental task in many fields. For example, in finance, an important task is to detect the fraud user activities in transaction data (Ahmed et al., 2016); in medical diagnosis, it is crucial to identify the patients whose symptoms/lab results are different from the normal population (Tarassenko et al., 1995; Cejnek et al., 2021); in proteomics, neuroscientists are interested in selecting the proteins that have higher levels of expression in the treatment condition compared with the negative controls (Shuster et al., 2022; Gao and Zhao, 2023). More generally, in some applications, the inliers (e.g., the non-fraud user activities) may follow different distributions in the reference and test datasets—this could happen if the reference dataset is collected with some preferences based on observed covariates. For example, one might include more individuals from the minority groups in the reference dataset to ensure representation. In these cases, the goal is then to distinguish the outliers from the inliers given an identifiable distribution shift in the inliers.

Formally, let $Z = (X, Y)$ denote a unit, where $X \in \mathcal{X}$ is the covariates and $Y \in \mathcal{Y}$ is the response. Assume that we are given a calibration dataset $\mathcal{D}_{\text{calib}} = \{Z_i\}_{i=1}^n$ that are assumed to be i.i.d. drawn from some distribution P. For a test dataset of independent units $\mathcal{D}_{\text{test}} = \{Z_{n+j}\}_{j=1}^m$, the objective is to decide whether each unit Z_{n+j} follows from Q, where

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(z) = w(x)$$

for some known weight function $w\colon \mathcal{X} \to \mathbb{R}^+$ (where $\frac{\mathrm{d}Q}{\mathrm{d}P}$ denotes the Radon-Nikodym derivative). Here, we are restrcting our attention to the distribution shift in the inliers that is entirely driven by the observable covariates. As a special case, if $w(z) \equiv 1$ (such that $Q = P$), we are back to the (vanilla) ourlier detection problem where the goal is to select test units that are different from the calibration units in distribution.

We consider this problem under the multiple testing framework, where we can define for each $j \in [m]$ the null hypothesis

$$H_j\colon Z_{n+j} \sim Q;$$

i.e., Z_{n+j} is an inlier. By rejecting H_j, we are expressing the belief that Z_{n+j} is an outlier in the dataset; the goal is to choose a subset of these hypotheses to reject while controlling the FDR.

This multiple testing problem has been studied in Bates et al. (2023) for $P = Q$, and generalized by Jin and Candès (2023) to allow for the covariate shift in the inliers. The methods proposed therein are based on the so-called (weighted) conformal p-values (Vovk et al., 2005). In what follows, we are to introduce the methods of conformalzed outlier detection, establish their equivalent e-BH interpretations, and then describe how to use conditional calibration to boost their power.

Throughout, we assume that we have access to a fixed nonconformity score function $V\colon \mathcal{Z} \to \mathbb{R}$ that assigns a score to each unit in $\{Z_i\colon i \in [n+m]\}$ such that a larger score translates to more evidence of being an outlier (though, it need not be accurate to guarantee FDR control). For example, V can be some model fitted on independent set-aside data.

6.1 Warm-up: conformal p-values and e-values

At its very simplest, conformalized outlier detection (with no covariate shift) uses $\mathcal{D}_{\text{calib}}$ and V to construct a p-value for each unit in $\mathcal{D}_{\text{test}}$. Using the shorthand $V_i := V(Z_i)$ for $i \in [n+m]$, the conformal p-value for H_j can be constructed as follows:

$$p_j = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i \geq V_{n+j}\}}{n+1}, \quad \forall j \in [m]. \tag{21}$$

It can be verified that under H_j, p_j is super-uniform, making it a valid p-value. In addition, under the alternative hypothesis, p_j is expected to be smaller—V_{n+j} should tend to be larger than the scores of the "conforming" calibration units. In the case of ties, we can add randomized tiebreakers to the numerator

in order to attain exact uniformity. Finally, Bates et al. (2023) show that although the conformal p-values are not independent, they are PRDS, so applying BH to the conformal p-values $(p_1, \ldots, p_m)$ will still control the FDR.

We now proceed to construct a collection of conformal e-values. First, define

$$T = \inf \left\{ t \in \{V_i\}_{i=1}^{n+m} : \frac{m}{n+1} \cdot \frac{1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \geq t\}}{(\sum_{j=1}^{m} \mathbb{1}\{V_{n+j} \geq t\}) \vee 1} \leq \alpha \right\}. \tag{22}$$

Using this threshold (which finds the smallest rejection threshold for the scores that controls an estimate of the FDR, similar to the model-X knockoff filter), we can construct a conformal e-value per hypothesis:

$$e_j = (n+1) \cdot \frac{\mathbb{1}\{V_{n+j} \geq T\}}{1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \geq T\}}, \quad \forall j \in [m]. \tag{23}$$

As we shall show later in Proposition 8, the evalue e_j defined above satisfies $\mathbb{E}[e_j] \leq 1$ for all $j \in \mathcal{H}_0$. Furthermore, applying e-BH to them will yield the same rejection set as applying BH to the conformal p-values defined in (21).

Remark 2. *Such an e-value construction (with a slight adjustment to the threshold T) was proposed in Bashari et al. (2024), whose authors use a martingale argument to prove that the conformal e-values are generalized e-values (Definition 1). By slightly modifying their proof strategy, we can show that the conformal e-values in the form of (23) are strict e-values, for a general class of thresholds including the one in (22) and the one in Bashari et al. (2024). The threshold in (22) is a special case that leads to the equivalence between BH and e-BH.*

Proposition 8. *Suppose $P = Q$. Given calibration data $\mathcal{D}_{\text{calib}}$ and test data $\mathcal{D}_{\text{test}}$, construct the conformal p-values $p_1, \ldots, p_m$ using (21) and conformal e-values $e_1, \ldots, e_m$ using (23). Let $\mathcal{R}^{\text{BH}} := \mathcal{R}^{\text{BH}}(p_1, \ldots, p_m)$ and $\mathcal{R}^{\text{e-BH}} := \mathcal{R}^{\text{e-BH}}(e_1, \ldots, e_m)$ be the rejection sets obtained by BH and e-BH, respectively, at some FDR control level $\alpha \in (0, 1)$. The following statements hold:*

(1) $\mathbb{E}[e_j] \leq 1$ for any $j \in \mathcal{H}_0$;

(2) $\mathcal{R}^{\text{BH}} = \mathcal{R}^{\text{e-BH}}$.

The proof of Proposition 8 is provided in Appendix A.5.

Remark 3. *Recall that one of the major sources e-BH's slackness is the inequality $\mathbb{1}\{e_j \geq \frac{m}{\alpha|\mathcal{R}|}\} \leq \alpha|\mathcal{R}|e_j/m$. With the threshold in (22), e-BH is almost tight in this step: e_j is either 0 or $(n+1)/(1 + \sum_{i\in[n]} \mathbb{1}\{V_i \geq T\})$; the nonzero value is very close to the e-BH threshold $m/(\alpha \sum_j \mathbb{1}\{V_{n+j} \geq T\})$ by the definition of T. This, however, is no longer the case when we extend this result to the weighted conformal e-values, or when one combines two e-values. In these cases, our conditional calibration method can be used to fill the gap and boost the power of e-BH.*

Remark 4. *As a side note, the selection set $\widehat{\mathcal{R}}^{\text{BH}}$ is also proposed by Gao and Zhao (2023) in a slightly different context. One might also notice that a third way to define the selection set is via $\widehat{\mathcal{R}}^{thres} = \{j \in [m] : V_{n+j} \geq T\}$, which is indenpendently proposed in Weinstein et al. (2017) and Mary and Roquain (2022) under different settings. It turns out this selection set $\widehat{\mathcal{R}}^{thres}$ is also equivalent to $\widehat{\mathcal{R}}^{\text{BH}}$ and $\widehat{\mathcal{R}}^{\text{e-BH}}$, which is implied by our proof of Proposition 8.*

6.2 Extension to covariate shift

We now return to the more general setting, where the inliers in $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{test}}$ are respectively drawn from distributions P and Q that can be different due to a covariate shift. Recall that we assume that the Radon-Nikodym derivative between P and Q is known to be $w(x)$—a function of only the covariates.

For such a setting, Tibshirani et al. (2019); Hu and Lei (2023); Jin and Candès (2023) have defined weighted analogues of the p-values in (21):

$$p_j = \frac{w(X_{n+j}) + \sum_{i=1}^{n} w(X_i)\mathbb{1}\{V_i \geq V_{n+j}\}}{w(X_{n+j}) + \sum_{i=1}^{n} w(X_i)}, \quad \forall j \in [m]. \tag{24}$$

It can be shown that p_j is still super-uniform and is thus valid. To achieve exact uniformity, we can again extend the construction by using random tie-breakers.

Jin and Candès (2023) show that the weighted conformal p-values no longer satisfy the PRDS condition, which means BH no longer guarantees FDR control when applied to weighted conformal p-values.

The authors propose a new procedure, called weighted conformal selection (WCS), which provably controls the FDR. We instead present a more straightforward e-BH alternative that is almost equivalent to (if not more powerful than) WCS. We will define weighted conformal e-values, which are weighted analogues of (23), and run e-BH. Similar to their unweighted versions, we first define the following hypothesis-specific thresholds: for $\forall j \in [m]$,

$$T_j = \inf\left\{t \in \{V_i\}_i^{n+m} \colon \frac{m}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)} \cdot \frac{w(X_{n+j}) + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq t\}}{(\sum_{k=1}^m \mathbb{1}\{V_{n+k} \geq t\}) \vee 1} \leq \alpha\right\}. \tag{25}$$

Using T_j, we then construct the e-value e_j:

$$e_j = \left(w(X_{n+j}) + \sum_{i=1}^n w(X_i)\right) \cdot \frac{\mathbb{1}\{V_{n+j} \geq T_j\}}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq T_j\}}, \quad \forall j \in [m]. \tag{26}$$

One can notice that in the presence of no covariate shift, where $w(x) \equiv 1$, the weighted conformal e-values coincide with their unweighted versions. The next proposition shows that the weighted conformal e-values defined above are still valid e-values; its proof is deferred to Appendix A.7.

Proposition 9. *For each $j \in [m]$, construct e_j by (26) with T defined in (25). Then $\forall j \in \mathcal{H}_0, \mathbb{E}[e_j] = 1$.*

By Proposition 9, applying e-BH to the weighted conformal e-values will control the FDR, but unlike the previous case, this is no longer equivalent to applying BH to the weighted conformal p-values.

The rejection set of e-BH, however, is (almost) identical to that of WCS with deterministic pruning. By "almost identical", we mean the following: Jin and Candès (2023) also propose an e-BH interpretation of WCS, and the corresponding e-values therein are provably no greater than the e-values constructed here (but the gap is usually very small). A detailed discussion on the connection will be delegated to the Appendix B. Effectively, we can expect the two methods to deliver similar empirical performance.

In practice, the power of WCS can be improved at the cost of randomization: instead of deterministically pruning to get the resulting rejection set, a randomized pruning rule can be used instead (Jin and Candès, 2023). The power gap between randomized and regular WCS can be quite significant in certain problem settings. This gap and the e-BH equivalence with WCS suggests that e-BH-CC has the potential to exhibit much higher power without invoking randomization. The next subsection demonstrates how we can fit the framework of conditional calibration to conformalized outlier detection.

6.3 Conditional calibration for weighted conformal selection

As in the other sections, we first identify the sufficient statistic for conditioning then prove that the constructed resampled e-values follow the desired conditional distribution. The proof of Proposition 9 suggests that the statistic $S_j = (\mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}})$ can serve as a conditioning statistic, where $\mathcal{E}_j$ is the unordered set of $\{Z_1, \ldots, Z_n, Z_{n+j}\}$ with repetitions allowed.

Formally, fix $j \in [m]$ and choose the sufficient statistic $S_j = (\mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}})$ of the combined data $\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}$. Conditional on S_j, resample

$$\widetilde{Z}_{n+j} \sim Z_{n+j} \mid \mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}} \sim \sum_{Z\in\mathcal{E}_j} \frac{w(X)}{\sum_{Z'\in\mathcal{E}_j} w(X')} \cdot \delta_Z, \tag{27}$$

where $Z = (X, Y)$, $Z' = (X', Y')$, and δ_a denotes the point mass at a. For the remaining elements $\{Z \in \mathcal{E}_j \colon Z \neq \widetilde{Z}_{n+j}\}$, arbitrarily assign them to $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$ without replacement by random permutation (their order will not matter—observe that T does not depend on the ordering of $\mathcal{D}_{\text{calib}}$). Using the resampled calibration dataset $\widetilde{\mathcal{D}}_{\text{calib}}^{(j)} = \{\widetilde{Z}_k\}_{k\in[n]}$ and resampled test dataset $\widetilde{\mathcal{D}}_{\text{test}}^{(j)} = \{\widetilde{Z}_{n+k}\}_{k\in[m]}$, where $\widetilde{Z}_{n+k} = Z_{n+k}$ for $k \neq j$, construct the thresholds (for each $k \in [m]$) defined in (25), denoted $\widetilde{T}_k^{(j)}$ to highlight its resampled nature. We then define the resampled e-values

$$\widetilde{e}_k^{(j)} = \left(w(\widetilde{X}_{n+j}) + \sum_{i=1}^n w(\widetilde{X}_i)\right) \cdot \frac{\mathbb{1}\{V(\widetilde{Z}_{n+k}) \geq \widetilde{T}_k^{(j)}\}}{w(\widetilde{X}_{n+k}) + \sum_{i=1}^n w(\widetilde{X}_i)\mathbb{1}\{V(\widetilde{Z}_i) \geq \widetilde{T}_k^{(j)}\}}, \quad \forall k \in [m], \tag{28}$$

where again we denote $\widetilde{Z}_i = (\widetilde{X}_i, \widetilde{Y}_i)$ for $i \in [n+m]$.

We express the correctness of our resampling scheme in the following proposition.

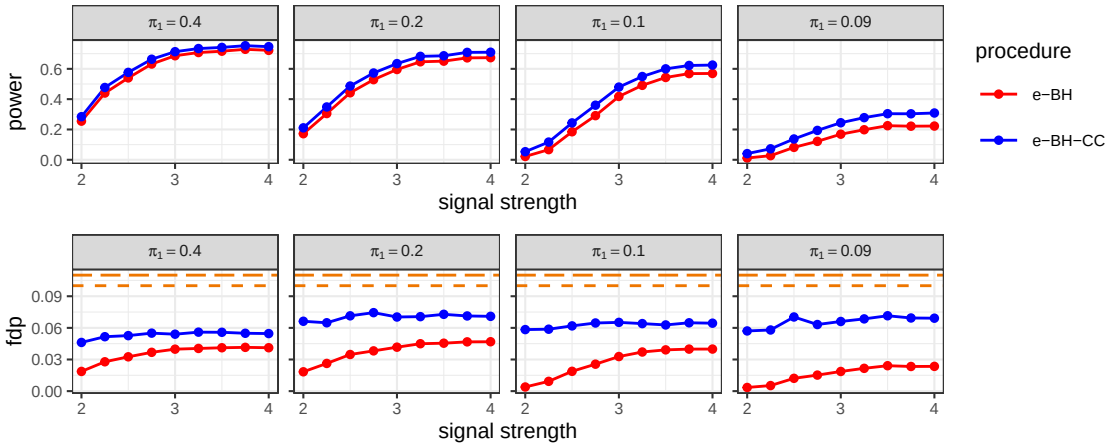

Figure 7: Realized power and FDP of the simulated experiments for unsupervised outlier detection. Each plot contains the averaged metrics over 1,000 replications. The orange dashed lines in the FDP plots represent the target FDR (0.1) and the FDR bound with Monte-Carlo error (0.11).

Proposition 10. *For each $j \in [m]$, choose sufficient statistic $S_j = (\mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}})$ and construct the thresholds $\widetilde{T}_k^{(j)}$ and their corresponding e-values $\widetilde{e}_k^{(j)}$ as defined in (28). Then*

$$\left(\widetilde{e}_1^{(j)}, \ldots, \widetilde{e}_m^{(j)}\right) \mid S_j \sim (e_1, \ldots, e_m) \mid S_j$$

under the null hypothesis $H_j\colon Z_{n+j} \sim Q$.

Furthermore, by choosing independent assignments of $(\widetilde{Z}_{n+j}, \widetilde{Z}_1, \ldots, \widetilde{Z}_n)$ by (27) for each resample, the corresponding resampled e-values are i.i.d. conditional on S_j.

The proof of Proposition 10 is provided in Appendix A.8.

6.4 Simulation studies

We run simulations in the unsupervised outlier detection problem (i.e., $Z_i = X_i$ for each $i \in [n+m]$) in the presence of covariate shift to show the power gained by using e-BH-CC on weighted conformal e-values (26).

Settings. We closely follow the outlier detection setup in Jin and Candès (2023). The authors state that as n, m get larger, the selection set of WCS approaches that of BH. To show the benefit of our method, we take $n = m = 200$ so that we avoid that limiting regime. In addition, to get a sense of the power drop and subsequent power gain over both sparse and dense outlier regimes, we vary the proportion of outliers $\pi_1 \in \{0.09, 0.1, 0.2, 0.4\}$.

At the beginning of each experiment, we sample 50 i.i.d. draws from $\text{Unif}([-3,3]^{50})$ to get an initial subset of points $\mathcal{W} \subseteq \mathbb{R}^{50}$. For each of the $\pi_1 n$ outliers in the test set, generate them i.i.d. as $X_{n+j} = \sqrt{a}L_{n+j} + W_{n+j}$, where $L_{n+j} \sim \mathcal{N}_{50}(0, I)$, $W_{n+j} \sim \text{Unif}(\mathcal{W})$ and the signal strength a varies through $\{2, 2.25, 2.5, \ldots, 4\}$. The inliers of the test set are instead generated i.i.d. as $X_{n+j} = L_{n+j} + W_{n+j}$, whose distribution we will denote as Q_X. The calibration dataset, all of which are inliers, is generated i.i.d from P_X, where $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x) = w(x) \propto \sigma(x^\top\theta)$. Here, $\sigma(\cdot)$ is the sigmoid function and $\theta \in \mathbb{R}^{50}$ is defined as $\theta_j = (0.3, 0.3, 0.2, 0.2, 0.1, 0.1, 0, \ldots, 0)$. As Jin and Candès (2023) state, this choice reflects a situation where the calibration dataset is sampled weighted by preference given by a logistic function of X.

Methods. We construct weighted conformal e-values as in (26), using $w(\cdot)$ as described above and $V(\cdot)$ equal to a one-class SVM trained on an independent hold-out set of 500 calibration inliers, weighted by $w(\cdot)$. We use the implementation found in the `scikit-learn` Python package, choosing the default `rbf` kernel. We then run e-BH and e-BH-CC, both at level $\alpha = 0.1$, on the resulting e-values.

Computational details. We again use a hybrid AVCS formulation to control the Monte-Carlo error (see Section 4.3 for details). The first 1,500 samples (in batches of 100) are used to construct the exact

AVCS, while the next 1,000 samples are used to construct the asymptotic AVCS. After 2,500 total samples, if the resulting AVCS still contains zero we stop early and fail to boost the particular e-value.

Experiments and results. For each choice of π_1 and a, we run 1,000 replications of both e-BH and e-BH-CC. Figure 7 reports their mean power and FDP over varying signal strengths for each choice of π_1. e-BH-CC is able to improve upon e-BH while continuing to control the FDR at α, regardless of the density of outliers. It is worth noting that there is a larger power gain when the outliers become sparser in the test dataset, reflecting the notion that the difficulty of the testing problem and the looseness of e-BH are interconnected.

7 Real data analysis

We apply our method to the national study of learning mindset (NSLM) observational dataset (Carvalho et al., 2019), with the goal of identifying individuals whose counterfactuals satisfy certain conditions. In this study, the intervention is to instill students with a learning mindset—the belief that intelligence can be developed, as opposed to being fixed—and the outcome of interest is their academic performance. The NSLM dataset contains observations on $n = 10{,}391$ students, among which 3,384 students received the intervention and 7,007 students did not. Each student is associated with a set of 11 covariates, including four student-level covariates and seven school-level covariates (more details about the covariates can be found in Table 1 of Carvalho et al. (2019)).

Problem setup. For each individual, denote by X the covariates, $T \in \{1, 0\}$ the intervention status (1 corresponds to receiving the intervention and 0 otherwise), and $Y \in \mathbb{R}$ the academic performance. To We adopt the potential outcome framework (Rubin, 1974), defining the potential outcomes $Y(1)$ and $Y(0)$ as the academic performance of a student with and without the intervention, respectively. Under the standard *stable unit treatment value assumption (SUTVA)* (Imbens and Rubin, 2015), the observed outcome $Y = TY(1)+(1-T)Y(0)$. Throughout, we also adopt the *strong ignorability assumption* (Imbens and Rubin, 2015), which states that

$$(Y(1), Y(0)) \perp\!\!\!\perp T \,|\, X.$$

Given the dataset $\{(X_i, T_i, Y_i)\}_{i\in[n]}$, we seek to detect the individuals with large $Y(1)$ in the control group, and those with small $Y(0)$ in the treatment group.

To proceed, we randomly split the dataset into three folds: the training fold $\mathcal{D}_{\text{train}}$, the calibration fold $\mathcal{D}_{\text{calib}}$, and the test fold $\mathcal{D}_{\text{test}}$. In what follows, we shall slightly abuse the notation and let $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, and $\mathcal{D}_{\text{test}}$ also refer to the corresponding index sets. Define $\mathcal{D}_{\text{calib}}(t) = \{i \in \mathcal{D}_{\text{calib}} : T_i = t\}$ and $\mathcal{D}_{\text{test}}(t) = \{i \in \mathcal{D}_{\text{test}} : T_i = t\}$ for $t = 0, 1$. We consider the following two tasks:

(1) *ATC-type*: find a subset of $\mathcal{D}_{\text{test}}(0)$ whose $Y(1) > 0.3$ with FDR controlled at level α.

(2) *ATT-type*: find a subset of $\mathcal{D}_{\text{test}}(1)$ with $Y(0) < -0.3$ with FDR controlled at level α.

In what follows, we focus on the ATC-type task, with the ATT-type task following by symmetry. Following Jin and Candès (2023), we consider this problem under the multiple testing framework, where the null hypothsis for any $j \in \mathcal{D}_{\text{test}}(0)$ is

$$H_j : Y_j(1) \le 0.3.$$

Here, rejecting H_j corresponds to identifying individual j who could have a large outcome. In slight contrast to the standard multiple testing setting, the null hypothesis H_j is random (see Jin and Candès (2023) for a detailed discussion) and we focus on the FDR defined as

$$\text{FDR}[\mathcal{R}] = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\{j \in \mathcal{R}\}}{|\mathcal{R}| \vee 1} \,\middle|\, \mathcal{H}_0\right].$$

The conformal e-values. For this problem, we adopt the data $\{i \in \mathcal{D}_{\text{calib}} : T_i = 1, Y_i \le 0.3\}$ as the calibration data. For the test points, those with $Y(1) \le 0.3$ are the inliers while the others are the outliers—our aim is to detect the outliers. Note that the inliers in the calibration set follow the distribution $P_{X,Y(1)\,|\,T=1,Y(1)\le 0.3}$, while the inliers in the test set follow the distribution $P_{X,Y(1)\,|\,T=0,Y(1)\le 0.3}$.

We can compute that the likelihood ratio is

$$\frac{dP_{X,Y(1)\,|\,T=0,Y(1)\le 0.3}}{dP_{X,Y(1)\,|\,T=1,Y(0)\le 0.3}} \propto \frac{1-e(x)}{e(x)}.$$

Above, $e(x) \coloneqq \mathbb{P}(T=1\,|\,X=x)$ is called the propensity score function. Since NSLM is an observational dataset, the propensity score function is unknown. We can nevertheless obtain an estimate of the propensity score function, $\widehat{w}(x)$, with the training fold $\mathcal{D}_{\text{train}}$ by regressing T on X. On $\mathcal{D}_{\text{train}}$, we can also determine the nonconformity score function $V(x,y)$ that is assumed to be non-increasing in y. With the fitted functions, Jin and Candès (2023) construct the conformal p-value for each $j \in \mathcal{D}_{\text{test}}(0)$ as in (24). Parallelly, we can define the conformal e-values as in (26), The result in Appendix B implies that applying e-BH to the conformal e-values is almost equivalent to—if not more powerful than—applying WCS to the conformal p-values.

Implementation and results. In our implementation, $|\mathcal{D}_{\text{train}}| = 8{,}000$, $|\mathcal{D}_{\text{calib}}| = 1{,}000$, and $|\mathcal{D}_{\text{test}}| = 1{,}000$. We use random forests to estimate the propensity score function $e(x)$ and the regression function $\widehat{m}(x)$ for $Y(1)\,|\,X=x$ with $\mathcal{D}_{\text{train}}$, where the random forests algorithm is implemented with the `scikit-learn` package in Python (Pedregosa et al., 2011). For each sample split, both WCS (its e-BH equivalent) with deterministic pruning and e-BH-CC are applied at FDR levels $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For reference, we also implement WCS with heterogeneous pruning—this is not meant to be compare with our method (since it is a randomized procudure), but rather to show the potential improvement over base WCS that e-BH-CC could achieve.

For e-BH-CC, we take α_{CC} to be a different value than α (recall Remark 1), where $\alpha_{\text{CC}} = 1.1 \cdot \alpha$. The AVCS-approximated e-values are adopted with $\alpha_0/\alpha = 0.0001$; the first 600 samples (in batches of 100) are used for constructing the non-asymptotic AVCS, and the next 400 for the asymptotic AVCS. If no conclusions can be drawn within 1000 Monte Carlo samples, we stop early and fail to boost the particular e-value. Figure 8 reports the average number of discoveries over 200 sample splits in the treated and control groups, respectively. We can see that e-BH-CC improves substantially over WCS with deterministic pruning, and is comparable to WCS with heterogeneous pruning, sending the message that using conditional calibration reclaims the power gap between the deterministic pruning and the randomized pruning.

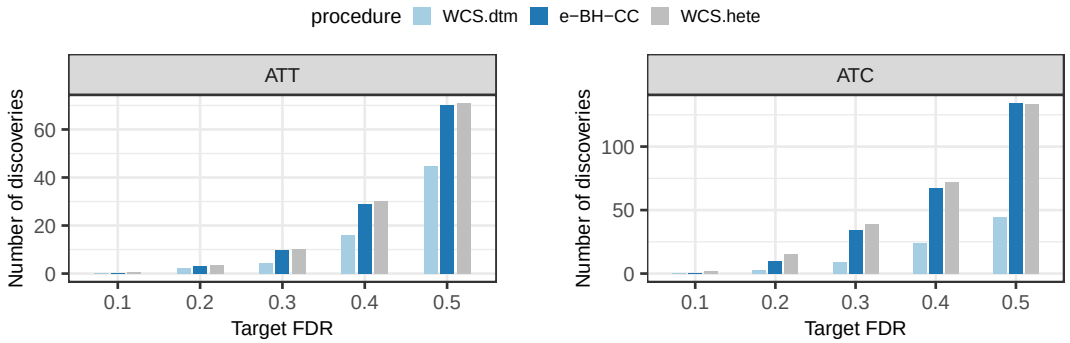

Figure 8: Number of identified students with positive treatment effects in the treatment group (left) and the control group (right) from the NSLM dataset. The results are averaged over 200 sample splits.

8 Discussion

In this paper, we introduce a framework to improve the power of e-BH via conditionally calibrating the e-values. Through three classes of multiple testing problems, we demonstrate how the proposed method can bring substantial power gains while continuing to control the FDR. We end this paper with a discussion on the potential applications of the proposed method, as well as future research directions.

Application to other selective inference problems Aside from the three examples discussed in this paper, we envision our framework to be useful for boosting many more selective inference procedures. For example, when testing m hypotheses with independent p-values $p_1, p_2, \ldots, p_m$, Li and Zhang (2023)

propose combining the rejection set of BH and SeqStep+ (Barber and Candès, 2015) through their e-value representations, and the resulting procedure is shown to be consistently better than the worse of BH and SeqStep+. For this procedure, our framework can immediately be applied by recognizing that $p_j \,|\, p_{-j} \sim \text{Unif}([0,1])$ under the null hypothesis. For another example, our framework can also be applied to other conformalized selection procedures (e.g., Bashari et al. (2024); Liang et al. (2022)) by considering similar sufficient statistics as introduced in Section 6. Lastly, we can even apply this framework to settings where p-values with arbtirary dependence are generated, as long as a viable sufficient statistic can be identfied. For example, Fithian and Lei (2022) describe problems such as edge testing in Gaussian graphical models and multiple comparisons with binary outcomes and identify possible sufficient statistics. By using a p-to-e calibrator (Vovk and Wang, 2021), we can construct e-values from the p-values and translate the problem into the e-BH framework. The p-to-e calibrator typically leads to power loss, but by using e-BH-CC we can likely regain the lost power.

Estimating the null proportion When the input e-values are strict e-values, e-BH as well as e-BH-CC controls the FDR at the level $\pi_0\alpha$, where π_0 is the fraction of null hypotheses. Such a guarantee can be too conservative, especially when the signal is sparse, i.e., $\pi_0 \ll 1$. One possible solution, borrowed from Fithian and Lei (2022), is to find an estimator $\widehat{\pi}_0$ of π_0, such that $\mathbb{E}[e_j/\widehat{\pi}_0] \le 1/\pi_0$, and then modify the ϕ_j function to be

$$\phi_j(c; S_j) = \mathbb{E}\left[\frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{ce_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\}}{|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} - \frac{e_j}{\widehat{\pi}_0} \,\middle|\, S_j\right].$$

We then define the boosted e-values and run the e-BH as in the current version. The challenge in this scheme boils down to finding the estimator $\widehat{\pi}_0$. It would be interesting to investigate the construction of $\widehat{\pi}_0$ in different problem settings.

Boosting via auxiliary statistics Our current boosting framework seeks for a multiplicative boosting factor $\widehat{c}_j$. It would be interesting to investigate other boosting forms to fill in the gap. For example, suppose for each $j \in [m]$, we have another statistic $W_j \in \mathbb{R}$ for testing H_j, where for simplicity, we assume that a large value of W_j suggest evidence against the null. We can then consider the "W_j-assisted" ϕ function:

$$\phi_j(c; S_j) = \mathbb{E}\left[\frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{e_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \text{ or } W_j \ge c\right\}}{|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} - e_j \,\middle|\, S_j\right].$$

Assuming we can (numerically) evaluate $\phi(c; S_j)$, we proceed to find the critical value $\widehat{c}_j$:

$$\widehat{c}_j := \sup\left\{c \in \mathbb{R} : \phi_j(c; S_j) \le 0\right\}.$$

A new collection of e-values boosted by W_j can be constructed as

$$e_j^{W,\mathfrak{b}} = \frac{m}{\alpha} \cdot \frac{\mathbb{1}\left\{e_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \text{ or } W_j \ge \widehat{c}_j\right\}}{|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}.$$

Since $\phi(+\infty; S_j) \le 0$, $\mathbb{E}[e_j^{W,\mathfrak{b}}] \le 1$ and $\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{W,\mathfrak{b}}) \supset \mathcal{R}(\boldsymbol{e})$. The proposed e-BH-CC is a specific instance of this formulation, where $W_j^{-1} = \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j$ and the threshold would be $\frac{1}{c}$. A future research direction is to investigate powerful choices of auxiliary statistics in different scenarios, and whether they may exhibit higher power than simply using e-BH-CC.

Acknowledgment

The authors would like to thank the Wharton Research Computing team for the computational resources provided and the great support from the staff members.

References

Ahmed, M., Mahmood, A. N., and Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288.

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

Bashari, M., Epstein, A., Romano, Y., and Sesia, M. (2024). Derandomized novelty detection with fdr control via conformal e-values. *Advances in Neural Information Processing Systems*, 36.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

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: series B (Methodological)*, 57(1):289–300.

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

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577.

Carvalho, C., Feller, A., Murray, J., Woody, S., and Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35.

Cejnek, M., Vysata, O., Valis, M., and Bukovsky, I. (2021). Novelty detection-based approach for alzheimer's disease and mild cognitive impairment diagnosis from eeg. *Medical & Biological Engineering & Computing*, 59:2287–2296.

Chi, Z., Ramdas, A., and Wang, R. (2022). Multiple testing under negative dependence. *arXiv preprint arXiv:2212.09706*.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics*, 34(2):275–297.

Ferreira, J. and Zwinderman, A. (2006). On the benjamini–hochberg method.

Fithian, W. and Lei, L. (2022). Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091 – 3118.

Gao, Z. and Zhao, Q. (2023). Simultaneous hypothesis testing using internal negative controls with an application to proteomics. *arXiv preprint arXiv:2303.01552*.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control.

Grünwald, P., de Heide, R., and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society, Series B (Methodology), with discussion*.

Grünwald, P. D. (2023). The e-posterior. *Philosophical Transactions of the Royal Society A*, 381(2247):20220146.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*.

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.

Hu, X. and Lei, J. (2023). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, pages 1–19.

Ignatiadis, N., Wang, R., and Ramdas, A. (2023). E-values as unnormalized weights in multiple testing. *Biometrika*, page asad057.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jin, Y. and Candès, E. J. (2023). Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*.

Larsson, M., Ramdas, A., and Ruf, J. (2024). The numeraire e-variable. *arXiv preprint arXiv:2402.18810*.

Li, G. and Zhang, X. (2023). E-values, multiple testing and beyond. *arXiv preprint arXiv:2312.02905*.

Liang, Z., Sesia, M., and Sun, W. (2022). Integrative conformal p-values for powerful out-of-distribution

testing with labeled outliers. *arXiv preprint arXiv:2208.11111*.

Luo, Y., Fithian, W., and Lei, L. (2022). Improving knockoffs with conditional calibration. *arXiv preprint arXiv:2208.09542*.

Mary, D. and Roquain, E. (2022). Semi-supervised multiple testing. *Electronic Journal of Statistics*, 16(2):4926–4981.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of markov's, chebyshev's and chernoff's inequalities. *arXiv preprint arXiv:2304.02611*.

Ren, Z. and Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sarkar, S. K. (2023). On controlling the false discovery rate in multiple testing of the means of correlated normals against two-sided alternatives. *arXiv preprint arXiv:2304.05261*.

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.

Shuster, S. A., Li, J., Chon, U., Sinantha-Hu, M. C., Luginbuhl, D. J., Udeshi, N. D., Carey, D. K., Takeo, Y. H., Xie, Q., Xu, C., et al. (2022). In situ cell-type-specific cell-surface proteomic profiling in mice. *Neuron*, 110(23):3882–3896.

Spector, A. and Janson, L. (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276.

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

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205.

Tarassenko, L., Hayton, P., Cerneaz, N., and Brady, M. (1995). Novelty detection for the identification of masses in mammograms.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Vladimir, V. and Wang, R. (2024). Nonparametric e-tests of symmetry. *New England Journal of Statistics in Data Science*.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

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

Vovk, V. and Wang, R. (2023). Confidence and discoveries with e-values. *Statistical Science*, 38(2):329–354.

Wang, Q., Wang, R., and Ziegel, J. (2022). E-backtesting. *arXiv preprint arXiv:2209.00991*.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

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

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., and Ramdas, A. (2021). Time-uniform central limit theory and asymptotic confidence sequences. *arXiv preprint arXiv:2103.06476*.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N., and Mineiro, P. (2022). Anytime-valid off-policy inference for contextual bandits. *ACM/JMS Journal of Data Science*.

Weinstein, A., Barber, R., and Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.

Xu, Z. and Ramdas, A. (2023a). More powerful multiple testing under dependence via randomization. *arXiv preprint arXiv:2305.11126*.

Xu, Z. and Ramdas, A. (2023b). Online multiple testing with e-values. *arXiv preprint arXiv:2311.06412*.

Xu, Z., Wang, R., and Ramdas, A. (2022). Post-selection inference for e-value based confidence intervals. *arXiv preprint arXiv:2203.12572*.

Zhang, Z., Ramdas, A., and Wang, R. (2023). On the existence of powerful p-values and e-values for composite hypotheses. *ArXiv preprint. arXiv:2305.16539*.

A Technical proofs

A.1 Proof of Theorem 1

Note that $\widehat{c}_j$ and $\phi_j(\widehat{c}_j; S_j)$ are both deterministic functions of S_j. For any $j \in \mathcal{H}_0$, we can write the expectation of $e_j^{\mathfrak{b}}$ as

$$\mathbb{E}[e_j^{\mathfrak{b}}] = \mathbb{E}\Big[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) \le 0\} \cdot \mathbb{E}[e_j^{\mathfrak{b}} \,|\, S_j] + \mathbb{1}\{\phi_j(\widehat{c}_j; S_j) > 0\} \cdot \mathbb{E}[e_j^{\mathfrak{b}} \,|\, S_j]\Big].$$

When $\phi_j(\widehat{c}_j; S_j) \le 0$, by construction of $e_j^{\mathfrak{b}}$ we have

$$\mathbb{E}[e_j^{\mathfrak{b}} \,|\, S_j] = \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{1}\left\{\widehat{c}_j e_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \,\middle|\, S_j\right] = \phi_j(\widehat{c}_j; S_j) + \mathbb{E}[e_j \,|\, S_j] \le \mathbb{E}[e_j \,|\, S_j].$$

When $\phi_j(\widehat{c}_j; S_j) > 0$, we let $\widehat{c}_{j,t} = \widehat{c}_j - 1/t$. Then we have determistically that

$$\mathbb{1}\big\{\widehat{c}_j \cdot e_j > \tfrac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\big\} = \lim_{t\to\infty} \mathbb{1}\big\{\widehat{c}_{j,t} \cdot e_j \ge \tfrac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\big\}.$$

As a result,

$$\begin{aligned}
\mathbb{E}[e_j^{\mathfrak{b}} \,|\, S_j] &= \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{1}\left\{\widehat{c}_j e_j > \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \,\middle|\, S_j\right] \\
&= \mathbb{E}\left[\lim_{t\to\infty} \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{1}\left\{\widehat{c}_{j,t} e_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \,\middle|\, S_j\right] \\
&\overset{\text{(a)}}{=} \lim_{t\to\infty} \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{1}\left\{\widehat{c}_{j,t} e_j \ge \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\} \,\middle|\, S_j\right] \\
&= \lim_{t\to\infty} \phi(\widehat{c}_{j,t}; S_j) \\
&\overset{\text{(b)}}{\le} \mathbb{E}[e_j \,|\, S_j].
\end{aligned}$$

Above, step (a) follows from the dominated covergence theorem and step (b) follows from the definition of the critical value $\widehat{c}_j$.

Combining the two cases, we have $\mathbb{E}[e_j^{\mathfrak{b}}] \le \mathbb{E}[e_j]$ for all $j \in \mathcal{H}_0$. Since $(e_1, \ldots, e_m)$ are valid (resp. generalized) e-values, the boosted e-values are valid (resp. generalized) e-values. The FDR control then follows from the FDR control of the e-BH procedure.

A.2 Proof of Proposition 1

For each $j \in \mathcal{H}_0$, we have

$$\mathbb{E}\big[e_j^{\mathfrak{b},\mathrm{CI}}\big] = \mathbb{E}\left[\frac{m\mathbb{1}\{U_{j,K} \le 0, \phi_j(\widetilde{c}_j; S_j) > 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] + \mathbb{E}\left[\frac{m\mathbb{1}\{U_{j,K} \le 0, \phi_j(\widetilde{c}_j; S_j) \le 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right].$$

Above, the first term can be bounded as follows:

$$\mathbb{E}\left[\frac{m\mathbb{1}\{\phi_j(\widetilde{c}_j; S_j) > U_{j,K}\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] = \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{P}\Big(\phi_j(\widetilde{c}_j; S_j) \notin C_{j,K} \,|\, S_j, \boldsymbol{e}\Big)\right] \le \mathbb{E}\left[\frac{m \cdot \alpha_{\mathrm{CI}}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] \le \alpha_0/\alpha.$$

As for the second term, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{m\mathbb{1}\{U_{j,K} \le 0, \phi_j(\widetilde{c}_j; S_j) \le 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] &\le \mathbb{E}\left[\frac{m\mathbb{1}\{\phi_j(\widetilde{c}_j; S_j) \le 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] \\
&= \underbrace{\mathbb{E}\left[\frac{m\mathbb{1}\{\widetilde{c}_j \le \widehat{c}_j, \phi_j(\widehat{c}_j; s_j) \le 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right]}_{\text{(i)}} + \underbrace{\mathbb{E}\left[\frac{m\mathbb{1}\{\widetilde{c}_j < \widehat{c}_j, \phi_j(\widehat{c}_j; s_j) > 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right]}_{\text{(ii)}}.
\end{aligned}$$

Recalling that $\widetilde{c}_j = \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}/e_j$, we then have

$$\begin{aligned}\text{(i)} = \mathbb{E}\left[\frac{m\mathbb{1}\{\widehat{c}_j e_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}, \phi_j(\widehat{c}_j; s_j) \leq 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] &= \mathbb{E}\Big[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) \leq 0\} \cdot \big(\phi_j(\widehat{c}_j; S_j) + \mathbb{E}[e_j \,|\, S_j]\big)\Big] \\ &\leq \mathbb{E}\big[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) \leq 0\} e_j\big].\end{aligned}$$

Next, we again take $\widehat{c}_{j,t} = \widehat{c}_j - 1/t$, and then we have

$$\begin{aligned}\text{(ii)} = \mathbb{E}\left[\frac{m\mathbb{1}\{\widehat{c}_j e_j > \frac{m}{\alpha|\widehat{\mathcal{R}}(\boldsymbol{e})|}, \phi_j(\widehat{c}_j; s_j) > 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] &= \mathbb{E}\left[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) > 0\} \cdot \mathbb{E}\left[\lim_{t\to\infty} \frac{m\mathbb{1}\{\widehat{c}_{j,t} e_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\}}{|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \,\middle|\, S_j\right]\right] \\ &\overset{\text{(a)}}{=} \lim_{t\to\infty} \mathbb{E}\left[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) > 0\} \cdot \mathbb{E}\left[\frac{m\mathbb{1}\{\widehat{c}_{j,t} e_j \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\}}{|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \,\middle|\, S_j\right]\right] \\ &= \lim_{t\to\infty} \mathbb{E}\Big[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) > 0\} \cdot \mathbb{E}\big[\phi_j(\widehat{c}_{j,t}) + e_j \,|\, S_j\big]\Big] \\ &\leq \mathbb{E}\big[\mathbb{1}\{\phi_j(\widehat{c}_j; S_j) > 0\} e_j\big],\end{aligned}$$

where step (a) applies the dominated convergence theorem. Combining (i) and (ii), we have bounded the second term by $\mathbb{E}[e_j]$, and therefore $\mathbb{E}[e_j^{\mathfrak{b},\mathrm{CI}}] \leq \mathbb{E}[e_j] + \alpha_0/\alpha$ for any $j \in \mathcal{H}_0$, invoking the proof of e-BH completes the proof.

Proof of Corollary 1 If $C_{j,K}$ is a $(1-\alpha_{\mathrm{CI}})$ asymptotic confidence interval for $\phi_j(\widehat{c}_j; S_j)$, it suffices to modify the upper bound of the first term in the proof of Proposition 1:

$$\begin{aligned}\lim_{K\to\infty} \mathbb{E}\left[\frac{m\mathbb{1}\{\phi_j(\widetilde{c}_j; S_j) > U_{j,K}\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] = \lim_{K\to\infty} \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\mathbb{P}\Big(\phi_j(\widetilde{c}_j; S_j) \notin C_{j,K} \,|\, S_j, \boldsymbol{e}\Big)\right] \\ = \mathbb{E}\left[\frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|} \lim_{K\to\infty} \mathbb{P}\Big(\phi_j(\widetilde{c}_j; S_j) \notin C_{j,K} \,|\, S_j, \boldsymbol{e}\Big)\right] \\ \leq \mathbb{E}\left[\frac{m \cdot \alpha_{\mathrm{CI}}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] \leq \alpha_0/\alpha.\end{aligned}$$

Again, the second step follows from the dominated convergence theorem.

A.3 Proof of Proposition 2

For any $j \in \mathcal{H}_0$, we have

$$\begin{aligned}\mathbb{E}[e_j^{\mathfrak{b},\mathrm{AVCS}}] &= \mathbb{E}\left[m \cdot \frac{\mathbb{1}\{\exists k, U_{j,k} \leq 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] \\ &= \mathbb{E}\left[m \cdot \frac{\mathbb{1}\{\exists k, U_{j,k} \leq 0, \phi_j(\widetilde{c}_j; S_j) > 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] + \mathbb{E}\left[m \cdot \frac{\mathbb{1}\{\exists k, U_{j,k} \leq 0, \phi_j(\widetilde{c}_j; S_j) \leq 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] \\ &\leq \mathbb{E}\left[m \cdot \frac{\mathbb{1}\{\exists k, U_{j,k} < \phi_j(\widetilde{c}_j; S_j)\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right] + \mathbb{E}\left[m \cdot \frac{\mathbb{1}\{\phi_j(\widetilde{c}_j; S_j) \leq 0\}}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right].\end{aligned}$$

Above, the first term is bounded by α_0/α by the construction of $\{C_{j,k}\}_{k\geq 1}$; the second term is bounded by $\mathbb{E}[e_j]$ following exactly the same steps in the proof of Proposition 1. Now that $\mathbb{E}[e_j^{\mathfrak{b},\mathrm{AVCS}}] \leq \mathbb{E}[e_j] + \alpha_0/\alpha$, invoking the proof of the e-BH procedure completes the proof.

Asymptotic anytime-valid confidence sequences The definition of an asymptotic anytime-valid confidence sequence below is adapted from Waudby-Smith et al. (2021).

Definition 2 (Asymptotic anytime-valid confidence sequences (Asymp-AVCS)). *We say that* $(\widehat{\theta}_k - L_k, \widehat{\theta}_k + U_k)_{k\geq 1}$ *centered around the estimators* $\{\widehat{\theta}_k\}_{k\geq 1}$ *with* $L_k, U_k > 0$ *for any* $k \geq 1$ *forms a* $(1-\alpha)$*-asymptotic anytime-valid confidence sequence for a parameter* θ *if there exists a non-asymptotic* $(1-\alpha)$*-anytime-valid confidence sequence* $(\widehat{\theta}_k - L_k^*, \widehat{\theta}_k + U_k^*)_{k\geq 1}$ *such that*

$$L_k^*/L_k \overset{a.s.}{\to} 1, \quad U_k^*/U_k \overset{a.s.}{\to} 1.$$

In practice, we can replace the $(1-\alpha_{\text{AVCS}})$-AVCS with an Asymp-AVCS when k is sufficiently large. The time-uniform coverage of the Asymp-AVCS has been established in Waudby-Smith et al. (2021) under certain conditions, and we refer the readers to their work for more details.

A.4 Proof of Proposition 3

We can write the intermediate e-values $e_j^{\mathfrak{b},\mathcal{S}} := e_j^{\mathfrak{b}} \mathbb{1}\{j \in \mathcal{S}\}$, for $j \in [m]$. Decomposing the FDR as in the proof of e-BH in (1), we see:

$$\begin{aligned}
\text{FDR}[\mathcal{S}] &= \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{j\in\mathcal{S}\}}{|\mathcal{S}|\vee 1}\right] = \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{j\in\mathcal{S}\}}{|\mathcal{S}\cup\{j\}|}\right] \\
&\overset{\text{(a)}}{\leq} \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{j\in\mathcal{S}\}\cdot\mathbb{1}\left\{e_j^{\mathfrak{b}} \geq \frac{m}{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}\right\}}{|\mathcal{R}(\boldsymbol{e})\cup\{j\}|}\right] \\
&\overset{\text{(b)}}{\leq} \sum_{j\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\{j\in\mathcal{S}\}\cdot\frac{\frac{\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|}{m}e_j^{\mathfrak{b}}}{|\mathcal{R}(\boldsymbol{e})\cup\{j\}|}\right] \\
&\overset{\text{(c)}}{=} \sum_{j\in\mathcal{H}_0} \frac{\alpha}{m}\mathbb{E}\left[\mathbb{1}\{j\in\mathcal{S}\}e_j^{\mathfrak{b}}\right] \\
&\overset{\text{(d)}}{\leq} \alpha.
\end{aligned}$$

Step (a) is due to $\mathcal{R}(\boldsymbol{e}) \subseteq \mathcal{S}$ (the denominator), and $\mathcal{S} \subseteq \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$: if $j \in \mathcal{S}$, we also have $j \in \mathcal{R}(\boldsymbol{e}^{\mathfrak{b}})$ and therefore $e_j^{\mathfrak{b}} = m/(\alpha|\widehat{\mathcal{R}}_j(\boldsymbol{e})|)$ by construction. Step (b) follows from the deterministic inequality $\mathbb{1}\{X \geq t\} \leq X/t$ for $t > 0$ in the e-BH proof. Step (c) is from the definition $\widehat{\mathcal{R}}_j(\boldsymbol{e}) = \mathcal{R}(\boldsymbol{e}) \cup \{j\}$. Step (d) follows since $e_j^{\mathfrak{b}}\mathbb{1}\{j \in \mathcal{S}\}$ is a valid e-value, as mentioned previously. Note that the entire inequality chain will also hold when $\boldsymbol{e}$ are generalized e-values.

A.5 Proof of Proposition 4

Let $\widetilde{\boldsymbol{e}} = (e_1^{\text{dBH}}/U_1, \ldots, e_m/U_m)$, and $\mathcal{R}^{\text{e-BH}}(\widetilde{\boldsymbol{e}})$ the selection set returned by the e-BH procedure applied to $\widetilde{\boldsymbol{e}}$. Let $\mathcal{R}^{\text{dBH}}(\boldsymbol{p})$ denote the selection set of the dBH procedure with U_j's. We denote by $\widetilde{e}_{(1)} \geq \widetilde{e}_{(2)} \geq \ldots \geq \widetilde{e}_{(m)}$ the ordered statistics of $\widetilde{\boldsymbol{e}}$ in descending order. Let $k^* = |\mathcal{R}^{\text{e-BH}}(\boldsymbol{e}^{\text{dBH}})|$, and recall that for dBH

$$r^* = \max\left\{r \in [m] : \left|\{j \in \widehat{\mathcal{R}}^+ : U_j \leq r/\widehat{R}_j(\boldsymbol{p})\}\right| \geq r\right\}$$

We can check that

$$\widetilde{e}_{(k)} \geq \frac{m}{\alpha k} \iff \left|\left\{j \in [m] : \widetilde{e}_j \geq \frac{m}{\alpha k}\right\}\right| \geq k.$$

As a result, by the definition of k^* and r^*,

$$\begin{aligned}
k^* = \max\left\{k \in [m] : \widetilde{e}_{(k)} \geq \frac{m}{\alpha k}\right\} &= \max\left\{k \in [m] : \left|\left\{j \in [m] : \widetilde{e}_j \geq \frac{m}{\alpha k}\right\}\right| \geq k\right\} \\
&= \max\left\{k \in [m] : \left|\left\{j \in \widehat{\mathcal{R}}^+ : \frac{m}{\alpha \widehat{R}_j(\boldsymbol{p}) U_j} \geq \frac{m}{\alpha k}\right\}\right| \geq k\right\} \\
&= \max\left\{k \in [m] : \left|\left\{j \in \widehat{\mathcal{R}}^+ : U_j \leq \frac{k}{\widehat{R}_j(\boldsymbol{p})}\right\}\right| \geq k\right\} \\
&= r^*,
\end{aligned}$$

where the second step follows from the construction of $\widetilde{e}_j$'s. This leads to

$$\begin{aligned}
\mathcal{R}^{\text{e-BH}}(\widetilde{\boldsymbol{e}}) = \left\{j \in [m] : \widetilde{e}_j \geq \frac{m}{\alpha k^*}\right\} &= \left\{j \in [m] : \frac{m\mathbb{1}\{j \in \widehat{\mathcal{R}}^+\}}{\alpha \widehat{R}_j(\boldsymbol{p}) U_j} \geq \frac{m}{\alpha k^*}\right\} \\
&= \left\{j \in \widehat{\mathcal{R}}^+ : \frac{r^*}{\widehat{R}_j(\boldsymbol{p})} \geq U_j\right\} \\
&= \mathcal{R}^{\text{dBH}}(\boldsymbol{p}).
\end{aligned}$$

The proof is therefore concluded.

A.6 Proof of Proposition 8

A.6.1 Proof of (a)

Define $\mathcal{V}$ to be the unordered collection of nonconformity scores $\{V_i : i \in [n+m]\}$. Let π be a permutation of $[n+m]$ such that $V_{\pi(1)} \leq \cdots \leq V_{\pi(n+m)}$, and let $N(k) := \sum_{i\in[n]} \mathbb{1}\{V_i \geq V_{\pi(k)}\}$, $R_0(k) = \sum_{j\in\mathcal{H}_0} \mathbb{1}\{V_{n+j} \geq V_{\pi(k)}\}$, and $R_1(k) = \sum_{j\in\mathcal{H}_1} \mathbb{1}\{V_{n+j} \geq V_{\pi(k)}\}$. Consider the discrete time filtration,

$$\mathcal{F}_k = \sigma\big(\mathcal{V}, \{N(\ell)\}_{\ell\leq k}, \{R_0(\ell)\}_{\ell\leq k}, \{R_1(\ell)\}_{\ell\leq k}\big), \text{ for } k \geq 1. \tag{29}$$

Instead of directly proving (a), we show the following stronger result. Suppose the conformal e-value takes the following form:

$$e_j = (n+1)\frac{\mathbb{1}\{V_{n+j} \geq T\}}{1 + \sum_{i\in[n]} \mathbb{1}\{V_i \geq T\}}, \text{ for } j \in [m]. \tag{30}$$

Then e_j is an e-value as long as $T = V_{\pi(\tau)}$, where is τ a stopping time adapted to the filtration $\{\mathcal{F}_k\}_{k\geq 1}$. Apparently, the threshold T defined in (22) satisfies the condition of Proposition 11, and therefore the e-value is a strict e-value.

The stronger result is stated in the following proposition, followed by its proof.

Proposition 11. *Under the setting of Proposition 8, let τ be a stopping time adapted to the filtration $\{\mathcal{F}_t\}_{k\geq 1}$ defined in (29). Suppose the e-value takes the form defined in (30). Then we have $\mathbb{E}[e_j] \leq 1$, for any $j \in \mathcal{H}_0$.*

Proof. To start, we define for $k \in \mathbb{N}_+$ that

$$M(k) = \frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\{V_{n+j} \geq V_{\pi(k)}\}}{1 + \sum_{i\in[n]} \mathbb{1}\{V_i \geq V_{\pi(k)}\}} = \frac{R_0(k)}{1 + N(k)},$$

where the last step is by the definition of $R_0(k)$ and $N(k)$. We claim that $M(k)$ is a supermartingale with respect to the filtration $\{\mathcal{F}_k\}_{k\geq 1}$. To see this, consider the following two scenarios.

(1) When $V_{\pi(k)} = V_{\pi(k-1)}$, we have $N(k) = N(k-1)$ and $R_0(k) = R_0(k-1)$, and thus $\mathbb{E}[M(k) \,|\, \mathcal{F}_{k-1}] = M(k-1)$.

(2) When $V_{\pi(k)} > V_{\pi(k-1)}$, we let $\Delta(k) = \sum_{i\in[n+m]} \mathbb{1}\{V_i = V_{\pi(k-1)}\}$, By the exchangeability of the inliers, conditional on $\mathcal{F}_{k-1}$, $N(k)$ follows a hypergeometric distribution with parameters $N(k-1)+R_0(k-1)$ (population size), $N(k-1)$ (success states in the population), and $N(k-1)+R_0(k-1)-\Delta(k)$ (number of draws). As a result, by direct computation of the conditional expectation with respect to the hypergeometric distribution (see e.g., Weinstein et al. (2017, Lemma 3.2)), we have $\mathbb{E}[M_k \,|\, \mathcal{F}_{k-1}] \leq M_{k-1}$.

Combing the two cases, we have shown that $M(k)$ is a supermartingale. Since τ is a stopping time with respect to $\{\mathcal{F}_k\}_{k\geq 1}$, applying the optional stopping theorem leads to

$$\mathbb{E}[M_\tau] \leq M_1 = \frac{|\mathcal{H}_0|}{1+n}. \tag{31}$$

Note also that

$$\mathbb{E}\left[\frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\{V_{n+j} \geq T\}}{\sum_{i\in[n]} \mathbb{1}\{V_i \geq T\}}\right] = \frac{1}{1+n}\sum_{j\in\mathcal{H}_0} \mathbb{E}[e_j] \tag{32}$$

Combining (31) and (32) yields $\sum_{j\in\mathcal{H}_0} \mathbb{E}[e_j] \leq |\mathcal{H}_0|$. Note also that τ is invariant to the permutation on $\{Z_{n+j}\}_{j\in\mathcal{H}_0}$, and thus by the exchangeability of $\{Z_{n+j}\}_{j\in\mathcal{H}_0}$, we have $\mathbb{E}[e_j] \leq 1$ for any $j \in \mathcal{H}_0$.

□

A.6.2 Proof of (2)

The statement is immediate from the following lemmas about T, the threshold defined in (22).

Lemma 1. $\widehat{\mathcal{R}}^{\mathrm{BH}} = \{j \in [m] \colon V_{n+j} \geq T\}$.

Proof. Let $V_{n+(1)} \geq V_{n+(2)} \geq \cdots \geq V_{n+(m)}$ denote the ordered nonconformity test scores in the descending order (with ties broken arbitrarily). By the definition of the conformal p-values, we also have $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

When $\widehat{\mathcal{R}}^{\mathrm{BH}}$ is empty, then for any $t \leq V_{n+(1)}$, let $m(t) = \sum_{j\in[m]} \mathbb{1}\{V_{n+j} \geq t\}$. By the definition of $m(t)$, one can check that $V_{n+(m(t))} \geq t$ and

$$\frac{m}{n+1}\frac{1+\sum_{i\in[n]}\mathbb{1}\{V_i \geq t\}}{\sum_{j\in[m]}\mathbb{1}\{V_{n+j}\geq t\}} = \frac{m}{n+1}\frac{1+\sum_{i\in[n]}\mathbb{1}\{V_i \geq t\}}{m(t)} = \frac{mp_{m(t)}}{m(t)} > \alpha.$$

The last step is because $\widehat{\mathcal{R}}^{\mathrm{BH}} = \varnothing$. By the definition of T, we have $T > V_{n+(1)}$, and thus $\widehat{\mathcal{R}}^{\mathrm{BH}} = \{j \in [m] : V_{n+j} \geq T\}$.

When $\widehat{\mathcal{R}}^{\mathrm{BH}}$ is nonempty, BH rejects the hypotheses corresponding to the $R \coloneqq |\widehat{\mathcal{R}}^{\mathrm{BH}}|$ largest nonconformity test scores, where

$$R = \max\left\{r \colon \frac{1+\sum_{i=1}^{n}\mathbb{1}\left\{V_i \geq V_{n+(r)}\right\}}{n+1} \leq \frac{\alpha r}{m}\right\}.$$

Observe that the rejection rule is equivalent to rejecting all nonconformity test scores that are at least

$$\widehat{v} = \max\left\{v \colon \frac{1+\sum_{i=1}^{n}\mathbb{1}\left\{V_i \geq v\right\}}{n+1} \leq \frac{\alpha \cdot |\{j \in [m] \colon V_{n+j} \geq v\}|}{m}\right\}.$$

We can rearrange the condition inside the maximum for the equivalent statement

$$\widehat{v} = \max\left\{v \colon \frac{m}{n+1}\cdot\frac{1+\sum_{i=1}^{n}\mathbb{1}\left\{V_i \geq v\right\}}{|\{j \in [m] \colon V_{n+j} \geq v\}|} \leq \alpha\right\}.$$

When the rejection set is non-empty, $\widehat{v}$ and T coincide. Thus, $\widehat{\mathcal{R}}^{\mathrm{BH}} = \{j \colon V_{n+j} \geq T\}$. □

Lemma 2. $\widehat{\mathcal{R}}^{\text{e-BH}} = \{j \in [m] \colon V_{n+j} \geq T\}$.

Proof. First, note that when $V_{n+j} < T$, the corresponding e-value is 0 and will never be rejected. Thus, we have $\widehat{\mathcal{R}}^{\text{e-BH}} \subseteq \{j \in [m] \colon V_{n+j} \geq T\}$. To prove the reverse inclusion, we can assume without loss of generality that $\{j \in [m] \colon V_{n+j} \geq T\}$ is nonempty (since otherwise the inclusion is trivial). Observe that when $V_{n+j} \geq T$, the e-value is positive and takes the value

$$e_j = \frac{n+1}{1+\sum_{i=1}^{n}\mathbb{1}\left\{V_i \geq T\right\}}$$

which does not depend on the index j (i.e., the nonzero e-values take the same value). Taking the definition of T directly, we see that

$$\frac{n+1}{1+\sum_{i=1}^{n}\mathbb{1}\left\{V_i \geq T\right\}} \geq \frac{m}{\alpha \cdot |\{j \in [n] \colon V_{n+j} \geq T\}|}.$$

However, $R \coloneqq |\{j \in [n] \colon V_{n+j} \geq T\}|$ is exactly the number of nonzero e-values in our collection. Each of the R nonzero e-values take the same value, which is bounded below by $\frac{m}{aR}$. By the e-BH procedure, each of these e-values will be selected, proving the reverse inclusion $\widehat{\mathcal{R}}^{\text{e-BH}} \supseteq \{j \colon V_{n+j} \geq T\}$. □

Since the two lemmas show equality of $\widehat{\mathcal{R}}^{\mathrm{BH}}$ and $\widehat{\mathcal{R}}^{\text{e-BH}}$ to the same rejection set, we conclude the desired proposition.

A.7 Proof of Proposition 9

For each $j \in [m]$, let us define an alternative threshold to (25):

$$\widehat{T}_j = \inf\left\{ t \in \{V_i\}_{i=1}^{n+m} : \frac{m}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)} \cdot \frac{w(X_{n+j})\mathbb{1}\{V_{n+j} \geq t\} + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq t\}}{1 + \sum_{k\in[m]\setminus\{j\}} \mathbb{1}\{V_{n+k} \geq t\}} \leq \alpha \right\}.$$

The significance of the above threshold is that on the event $\{V_{n+j} \geq T_j\}$, $T_j = \widehat{T}_j$. To see this, we first note that by construction, $\widehat{T}_j \leq T_j$. On the event $\{V_{n+j} \geq T_j\}$, there is also $V_{n+j} \geq \widehat{T}_j$. To prove the inverse inequality, consider the following quantity

$$\begin{aligned}
&\frac{m}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)} \cdot \frac{w(X_{n+j}) + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq \widehat{T}_j\}}{\mathbb{1}\{V_{n+j} \geq \widehat{T}_j\} + \sum_{k\in[m]\setminus\{j\}} \mathbb{1}\{V_{n+k} \geq \widehat{T}_j\}} \\
= &\frac{m}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)} \cdot \frac{w(X_{n+j})\mathbb{1}\{V_{n+j} \geq \widehat{T}_j\} + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq \widehat{T}_j\}}{1 + \sum_{k\in[m]\setminus\{j\}} \mathbb{1}\{V_{n+k} \geq \widehat{T}_j\}} \leq \alpha,
\end{aligned}$$

where the first step is because $V_{n+j} \geq T_j \geq \widehat{T}_j$ and the second step is by the definition of $\widehat{T}_j$. We can then conclude that $T_j \leq \widehat{T}_j$, and therefore $T_j = \widehat{T}_j$.

In addition, define the set $\mathcal{E}_j$ to be the unordered collection of $\{Z_1, Z_2, \ldots, Z_n, Z_{n+j}\}$ (with repetitions allowed). Then $\widehat{T}_j$, formally constructed using each of $Z_1, \ldots, Z_{n+m}$, actually only depends on the data through $\mathcal{E}_j$ and $\{Z_{n+k}\}_{k\in[m]\setminus\{j\}}$. That is, $\widehat{T}_j$ is agnostic to the ordering of the elements in $\mathcal{E}_j$.

Using these facts, we can analyze the expectation of e_j under the null H_j:

$$\begin{aligned}
\mathbb{E}[e_j] &= \mathbb{E}\left[\frac{\big(w(X_{n+j}) + \sum_{i=1}^n w(X_i)\big)\mathbb{1}\{V_{n+j} \geq T_j\}}{w(X_{n+j}) + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq T_j\}}\right] \\
&= \mathbb{E}\left[\frac{\big(w(X_{n+j}) + \sum_{i=1}^n w(X_i)\big)\mathbb{1}\{V_{n+j} \geq T_j\}}{w(X_{n+j})\mathbb{1}\{V_{n+j} \geq T_j\} + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq T_j\}}\right] \\
&= \mathbb{E}\left[\frac{\big(w(X_{n+j}) + \sum_{i=1}^n w(X_i)\big)\mathbb{1}\{V_{n+j} \geq \widehat{T}_j\}}{w(X_{n+j})\mathbb{1}\{V_{n+j} \geq \widehat{T}_j\} + \sum_{i=1}^n w(X_i)\mathbb{1}\{V_i \geq \widehat{T}_j\}}\right] \\
&= \mathbb{E}\Bigg[\frac{w(X_{n+j}) + \sum_{i=1}^n w(X_i)}{w(X_{n+j})\mathbb{1}\{V_{n+j} \geq \widehat{T}_j\} + \sum_{i=1}^n w(Z_i)\mathbb{1}\{V_i \geq \widehat{T}_j\}} \\
&\qquad \times \mathbb{E}\Big[\mathbb{1}\Big\{V_{n+j} \geq \widehat{T}_j\Big\} \,\Big|\, \mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}}\Big]\Bigg].
\end{aligned} \tag{33}$$

The second equality technically adopts the notation $0/0 = 0$. The third equality follows from the equivalence of T_j and $\widehat{T}_j$ on the event $\mathbb{1}\{V_{n+j} \geq T_j\} = 1$ as discussed above. The last step uses the tower property of conditional expectation.

Under the null H_j, we have following characterization of the conditional distribution of Z_{n+j} given $\mathcal{E}_j$ (Jin and Candès, 2023):

$$Z_{n+j} \mid \{\mathcal{E}_j = z\} \sim \sum_{k\in[n]\cup\{n+j\}} \frac{w(z_k)}{\sum_{i=1}^n w(z_i) + w(z_{n+j})} \cdot \delta_{z_k}, \tag{34}$$

where δ_a denotes a point mass at a. In the above, $z = \{z_1, z_2, \ldots, z_n, z_{n+j}\}$ is a realization of $\mathcal{E}_j$, with $z_i = (x_i, y_i)$. From (34), it is immediate that for a constant t (conditional on $\mathcal{E}_j$), $\mathbb{P}(V_{n+j} \geq t \mid \mathcal{E}_j)$ is equal to the weighted sum of indicator random variables:

$$\mathbb{P}(V_{n+j} \geq t \mid \mathcal{E}_j) = \sum_{k\in[n]\cup\{n+j\}} \frac{w(x_k)\mathbb{1}\{V(z_k) \geq t\}}{w(x_{n+j}) + \sum_{i=1}^n w(x_i)}. \tag{35}$$

Since

$$\mathbb{E}\Big[\mathbb{1}\Big\{V_{n+j} \geq \widehat{T}_j\Big\} \mid \mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}}\Big] = \mathbb{P}\big(V_{n+j} \geq \widehat{T}_j \mid \mathcal{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}}\big),$$

$\widehat{T}_j$ is constant conditioned on $\mathcal{E}_j$ and $\{Z_{n+k}\}_{k\in[m]\setminus\{j\}}$, and V_{n+j} is independent of $\{Z_{n+k}\}_{k\in[m]\setminus\{j\}}$ by assumption, we can use (35) to directly conclude that

$$\mathbb{E}\left[\mathbb{1}\left\{V_{n+j}\geq\widehat{T}_j\right\}\mid\mathcal{E}_j,\{Z_{n+k}\}_{k\in[m]\setminus\{j\}}\right]=\frac{w(X_{n+j})\mathbb{1}\{V_{n+j}\geq\widehat{T}_j\}+\sum_{i=1}^n w(X_i)\mathbb{1}\{V_i\geq\widehat{T}_j\}}{w(X_{n+j})+\sum_{i=1}^n w(X_i)}.$$

The above in conjunction with (33) implies that $\mathbb{E}[e_j]=1$.

A.8 Proof of Proposition 10

The proposition is immediate if we can conclude that $\widetilde{Z}_1,\widetilde{Z}_2,\ldots,\widetilde{Z}_{n+m}$ and $Z_1,Z_2,\ldots,Z_{n+m}$ are jointly equal in distribution conditioned on $(\mathcal{E}_j,\{Z_{n+k}\}_{k\in[m]\setminus\{j\}})$. Since $\widetilde{Z}_{n+k}=Z_{n+k}$ for $k\neq j$ by construction, we only need to consider the joint distribution corresponding to the indices $[n]\cup\{n+j\}$ (conditional on $\mathcal{E}_j$, as the other test units were independently drawn).

Assuming the null H_j, we can write joint probability density function of $Z_1,\ldots,Z_n,Z_{n+j}$ in terms of the weight function w and the density function p of P (Tibshirani et al., 2019):

$$f(z_1,z_2,\ldots,z_n,z_{n+j})=w(x_{n+j})\prod_{i\in[n]\cup\{n+j\}}p(z_i).^{4}$$

We can use this to calculate the joint conditional probabilities. Treat $\mathcal{E}_j$ as fixed and denote its elements as $\{z_1,z_2,\ldots,z_n,z_{n+1}\}$, without any particular order. Let $\mathcal{S}_{n+1}$ denote all permutations of $[n+1]$. Then for any permutation $\sigma\in\mathcal{S}_{n+1}$, we have

$$\begin{aligned}\mathbb{P}(Z_{n+j}=z_{\sigma(n+1)},Z_1=z_{\sigma(1)},\ldots,Z_n=z_{\sigma(n)}\mid\mathcal{E}_j)&=\frac{w(x_{\sigma(n+1)})\prod_{i=1}^{n+1}p(z_i)}{\sum_{\sigma'\in\mathcal{S}_{n+1}}w(x_{\sigma'(n+1)})\prod_{i=1}^{n+1}p(z_i)}\\&=\frac{w(x_{\sigma(n+1)})}{\sum_{i=1}^{n+1}w(x_i)\cdot n!}.\end{aligned}$$

Meanwhile, the event $\{\widetilde{Z}_{n+j}=z_{\sigma(n+1)},\widetilde{Z}_1=z_{\sigma(1)},\ldots,\widetilde{Z}_n=z_{\sigma(n)}\}\mid\mathcal{E}_j$ occurs when we first resample $\widetilde{Z}_{n+j}$ to be $z_{\sigma(n+1)}$ and subsequently assign $\{z_1,z_2,\ldots,z_{n+1}\}\setminus\{z_{\sigma(n+1)}\}$ to $\widetilde{Z}_1,\ldots,\widetilde{Z}_n$ uniformly at random (without replacement). Thus,

$$\mathbb{P}(\widetilde{Z}_{n+j}=z_{\sigma(n+1)},\widetilde{Z}_1=z_{\sigma(1)},\ldots,\widetilde{Z}_n=z_{\sigma(n)}\mid\mathcal{E}_j)=\frac{w(x_{\sigma(n+1)})}{\sum_{i=1}^{n+1}w(x_i)}\cdot\frac{1}{n!}.$$

We conclude that the conditional joint distributions match, as desired. The i.i.d. property of the resamples follows immediately by observing that the only randomness of the e-values, conditional on S_j, comes from the random (weighted) assignment to $(\widetilde{Z}_{n+j},\widetilde{Z}_1,\ldots,\widetilde{Z}_n)$.

B Connection between eBH and WCS

For the outlier detation problem, the WCS procedure (Jin and Candès, 2023) with deterministic pruning computes a weighted conformal p-value for each $j\in[m]$ as in (24) and returns the selection set

$$\mathcal{R}^{\mathrm{WCS}}=\left\{j\in[m]:p_j\leq\frac{\alpha}{m}\cdot|\widehat{\mathcal{R}}_j|,\ |\widehat{\mathcal{R}}_j|\leq r^*\right\}.$$

Above, $\widehat{\mathcal{R}}_j$ is a "proxy" selection set, obtained via applying the BH procedure to $(p_1^{(j)},\ldots,p_{j-1}^{(j)},0,p_{j+1}^{(j)},\ldots,p_m^{(j)})$, where for each $\ell\in[m]$,

$$p_\ell^{(j)}=\frac{\sum_{i\in[n]}w(X_i)\cdot\mathbb{1}\{V_i\geq V_{n+\ell}\}+w(X_{n+j})\cdot\mathbb{1}\{V_{n+j}\geq V_{n+\ell}\}}{\sum_{i\in[n]}w(X_i)+w(X_{n+j})},$$

and the threshold r^* is defined via

$$r^*=\max\left\{r\in[m]:\sum_{j\in[m]}\mathbb{1}\left\{p_j\leq\alpha|\widehat{R}_j|/m,\ |\widehat{R}_j|\leq r\right\}\geq r\right\}.$$

[4] As in Tibshirani et al. (2019), we use the term "density function" loosely to refer to the Radon-Nikodym derivative with respect to an arbitrary base measure.

As pointed out by Jin and Candès (2023), $\mathcal{R}^{\mathrm{WCS}}$ can equivalently obtained by applying the BH procedure to the e-values:

$$e_j^{\mathrm{WCS}} = \frac{\mathbb{1}\{p_j \le \alpha|\widehat{R}_j|/m\}}{\alpha|\widehat{R}_j|/m}, \quad \forall j \in [m].$$

We are about to show that the weighted conformal e-value e_j constructed in (26) satisfies $e_j \ge e_j^{\mathrm{WCS}}$ determistically. The following lemma is key to establish this connection.

Lemma 3. *For any $j \in [m]$, the following holds.*

(1) On the event that $\{p_j \le \alpha|\widehat{R}_j|/m\}$, $\widehat{R}_j = \{\ell \in [m] \colon V_{n+\ell} \ge T_j\}$.

(2) $p_j \le \alpha|\widehat{R}_j|/m$ if and only if $V_{n+j} \ge T_j$;

Lemma 3 then implies that

$$\begin{aligned} e_j &= \Big(w(X_{n+j}) + \sum_{i=1}^{n} w(X_i)\Big) \cdot \frac{\mathbb{1}\{V_{n+j} \ge T_j\}}{\sum_{i\in n} w(X_i)\mathbb{1}\{V_i \ge T_j\} + w(X_{n+j})} \\ &\overset{\text{(a)}}{\ge} \frac{m}{\alpha} \cdot \frac{\mathbb{1}\{V_{n+j} \ge T_j\}}{|\{\ell \in [m] \colon V_{n+\ell} \ge T_j\}|} \overset{\text{(b)}}{=} \frac{m}{\alpha} \cdot \frac{\mathbb{1}\{p_j \le \alpha|\widehat{R}_j|/m\}}{\widehat{R}_j} = e_j^{\mathrm{WCS}}. \end{aligned}$$

Above, step (a) follows from the definition of T_j and step (b) uses Lemma 3. We also note that step (a) is often quite tight due to the choice of T_j.

Now, the only missing piece is the proof of Lemma 3, which we provide below.

Proof of Lemma 3.

(1) Fix $j \in [m]$, and let $k^* = |\widehat{R}_j|$ and $\ell^* = |\{\ell \in [m] : V_{n+\ell} \ge T_j\}|$.

Suppose $p_j \le \alpha|\widehat{R}_j|/m$. By the property of the BH procedure,

$$\widehat{R}_j = \{j\} \cup \{\ell \ne j : p_\ell^{(j)} \le \alpha k^*/m\} = \{\ell \in [m] : p_\ell^{(j)} \le \alpha k^*/m\},$$

where the last step is because $p_j^{(j)} = p_j \le \alpha k^*/m$. Let $p_{(1)}^{(j)} \le \cdots \le p_{(m)}^{(j)}$ denote the ordered p-values in an ascending order. Since the rank of $p_\ell^{(j)}$ is determined by $V_{n+\ell}$, we have (with a slight abuse of notation) that $V_{n+(1)} \ge V_{n+(2)} \ge \cdots \ge V_{n+(m)}$.

Next, note that

$$\frac{\sum_{i\in[n]} w(X_i)\mathbb{1}\{V_i \ge V_{n+(k^*)}\} + w(X_{n+j})}{\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+(k^*)}\}} \frac{m}{w(X_{n+j}) + \sum_{i\in[n]} w(X_i)} = p_{(k^*)}^{(j)} \frac{m}{k^*} \le \alpha. \tag{36}$$

where the first inequality follows from the definition of $p_{k^*}^{(j)}$ and that $\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+(k^*)}\} = k^*$; the last step is due to the property of the BH procedure. Eqn. (36) implies that $V_{n+(k^*)} \ge T_j$. Therefore for any $\ell \in \widehat{R}_j$, $V_{n+\ell} \ge V_{n+(k^*)} \ge T_j$, and $\widehat{R}_j \subset \{\ell \in [m] : V_{n+\ell} \ge T_j\}$.

Conversely, we can see that

$$\begin{aligned} p_{(\ell^*)}^{(j)} &= \frac{\sum_{i\in[n]} w(X_i)\mathbb{1}\{V_i \ge V_{n+(\ell^*)}\} + w(X_{n+j})}{\sum_{i\in[n]} w(X_i) + w(X_{n+j})} \\ &\overset{\text{(a)}}{=} \frac{\sum_{i\in[n]} w(X_i)\mathbb{1}\{V_i \ge V_{n+(\ell^*)}\} + w(X_{n+j})}{\sum_{\ell\in m} \mathbb{1}\{V_{n+\ell} \ge V_{n+(\ell^*)}\}} \frac{m}{\sum_{i\in[n]} w(X_i) + w(X_{n+j})} \frac{\ell^*}{m} \\ &\overset{\text{(b)}}{\le} \frac{\alpha\ell^*}{m}. \end{aligned} \tag{37}$$

Above, step (a) is because $\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+(\ell^*)}\} = \ell^*$, and step (b) is because $V_{n+(\ell^*)} \ge T_j$. As a result of (37) and the property of the BH procedure, we have $\ell^* \le k^*$. For any $\ell \in [m]$ such that $V_{n+\ell} \ge T_{n+j}$, there is also $p_\ell \le p_{(\ell^*)} \le \alpha k^*/m$, implying that $\ell \in \widehat{R}_j$. Collectively, we have $\widehat{R}_j = \{\ell \in [m] : V_{n+\ell} \ge T_j\}$.

(2) When $p_j \le \alpha|\widehat{R}_j|/m$, we have by (1) that $p_j \le \frac{\alpha\ell^*}{m}$. Suppose otherwise $V_{n+j} < T_j$. We can check that

$$\begin{aligned}
&\frac{\sum_{i\in[n]} w(X_i)\mathbb{1}\{V_i \ge V_{n+j}\} + w(X_{n+j})}{\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+j}\}} \frac{m}{\sum_{i\in[n]} w(X_i) + w(X_{n+j})} \\
&= p_j \cdot \frac{m}{\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+j}\}} \\
&\le p_j \cdot \frac{m}{\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge T_j\}} \\
&= \frac{p_j m}{\ell^*} \le \alpha.
\end{aligned}$$

The above implies that $V_{n+j} \ge T_j$, which is a contradiction. Therefore, we conclude that $V_{n+j} \ge T_j$. Conversely, when $V_{n+j} \ge T_j$, there is

$$\frac{\sum_{i\in[n]} w(X_i)\mathbb{1}\{V_i \ge V_{n+j}\} + w(X_{n+j})}{\sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+j}\}} \frac{m}{\sum_{i\in[n]} w(X_i) + w(X_{n+j})} \le \alpha.$$

Rearranging the above inequality, we have

$$p_j \le \frac{\alpha}{m} \sum_{\ell\in[m]} \mathbb{1}\{V_{n+\ell} \ge V_{n+j}\} \le \frac{\alpha\ell^*}{m}.$$

Recall that Eqn. (37) proves $p^{(j)}_{(\ell^*)} \le \alpha\ell^*/m$. By the property of the BH procedure, $\ell^* \le |\mathcal{R}^{\mathrm{BH}}(p^{(j)}_1, \ldots, p^{(j)}_m)| \le |\widehat{R}_j|$. As a result, we have $p_j \le \alpha|\widehat{R}_j|/m$, completing the proof.