False Discovery Rate Adjustments for Average Significance Level Controlling Tests

Timothy B. Armstrong*
University of Southern California

June 11, 2026

Abstract

Multiple testing adjustments, such as the Benjamini and Hochberg (1995) step-up procedure for controlling the false discovery rate (FDR), are typically applied to families of tests that control significance level in the classical sense: for each individual test, the probability of false rejection is no greater than the nominal level. In this paper, we consider tests that satisfy only a weaker notion of significance level control, in which the probability of false rejection need only be controlled on average over the hypotheses. We find that the Benjamini and Hochberg (1995) step-up procedure still controls FDR in the asymptotic regime with many weakly dependent p-values and an increasing number of rejections, and that certain adjustments for dependent p-values such as the Benjamini and Yekutieli (2001) procedure continue to yield FDR control in finite samples. Our results open the door to FDR controlling procedures in nonparametric and high dimensional settings where weakening the notion of inference may allow for power improvements.

1 Introduction

Consider testing m hypotheses $H_1, \ldots, H_m$. Let $\mathcal{H}_0 \subseteq \{1, \ldots, m\}$ denote the set of true null hypotheses. Given p-values $p_1, \ldots, p_m$ for each of the hypotheses, we wish to form a multiple testing procedure which decides on a subset of hypotheses to reject. A common

*email: timothy.armstrong@usc.edu. Support from National Science Foundation Grant SES-2049765 is gratefully acknowledged.

starting point for multiple testing procedures proposed in the literature is to assume that the p-values are formed from tests that control significance level in the classical sense, which implies

$$\text{for all } t \in [0,1] \text{ and } i \in \mathcal{H}_0,\ P(p_i \leq t) \leq t. \tag{1}$$

One then adjusts the critical value so that some notion of multiple testing error, such as the false discovery rate (FDR), is controlled (see formal definitions below).

In this paper, we explore the possibility of forming FDR controlling multiple testing procedures from tests that satisfy a weaker *average significance level* control criterion:

$$\text{for all } t \in [0,1],\ \frac{1}{m}\sum_{i \in \mathcal{H}_0} P(p_i \leq t) \leq t. \tag{2}$$

Such tests can be formed from confidence intervals (CIs) that weaken the classical definition of a CI by requiring coverage only on average over the reported CIs. Letting $CI_1(t), \ldots, CI_m(t)$ be CIs for parameters $\theta_1, \ldots, \theta_m$ with nominal $100 \cdot (1-t)\%$ coverage, the *average coverage* criterion requires

$$\frac{1}{m}\sum_{i=1}^{m} P(\theta_i \notin CI_i(t)) \leq t. \tag{3}$$

Given null hypotheses $H_i : \theta_i = \theta_{0,i}$, p-values formed from these CIs will, by definition, satisfy $p_i \leq t$ iff. $\theta_{0,i} \notin CI_i(t)$. If the CIs satisfy (3) for each $t \in [0,1]$, the resulting p-values will satisfy (2) since $\frac{1}{m}\sum_{i \in \mathcal{H}_0} P(\theta_{0,i} \notin CI_i(t)) = \frac{1}{m}\sum_{i \in \mathcal{H}_0} P(\theta_i \notin CI_i(t)) \leq \frac{1}{m}\sum_{i=1}^{m} P(\theta_i \notin CI_i(t)) \leq t$.

CIs satisfying the average coverage criterion (3) and related criteria have been developed in a number of settings (Wahba, 1983; Nychka, 1988; Wasserman, 2007, Chapter 5.8; Cai et al., 2014; Armstrong et al., 2022). They are particularly appealing in high dimensional or nonparametric settings involving regularized estimation, where impossibility results (Low, 1997) severely restrict the scope for constructing classical tests and CIs. Additional settings where the average significance level condition (2) can be shown to hold have been considered in recent work by Ignatiadis and Sen (2025), Ignatiadis et al. (2025) and Barber and Samworth (2025).

We ask whether p-values satisfying the weaker condition (2) can be used as an input to multiple testing procedures used in the literature. We focus on multiple testing procedures

designed to control the false discovery rate (FDR) of Benjamini and Hochberg (1995). We find that average significance level control is indeed sufficient for certain multiple testing procedures to guarantee FDR control. In particular, average significance level control is sufficient to guarantee FDR control of the Benjamini and Hochberg (1995) procedure in the asymptotic regime of weakly dependent p-values and many hypotheses ($m \to \infty$) and of the Benjamini and Yekutieli (2001) procedure with fixed m and arbitrary dependence among p-values. On the other hand, in contrast to the classical setting, we show by example that the Benjamini and Hochberg (1995) procedure does not in general have FDR control with fixed m and independent p-values, and that approaches that estimate the proportion of null hypotheses, such as the procedure of Storey (2002), can fail to control FDR even as $m \to \infty$.

Much of the literature on FDR controlling multiple testing procedures takes a family of p-values satisfying the classical significance level control condition (1) as a starting point. An important exception is the literature on knockoff based FDR controlling procedures (Barber and Candès, 2015), which instead rely on the construction of auxiliary random variables, called knockoffs. Constructing knockoffs typically requires modeling assumptions such as the "model-X" framework, in which the joint distribution of regression covariates is known or estimated with sufficient accuracy (Candès et al., 2018), or restricting the procedure to low dimensional settings; see also Arias-Castro and Chen (2017) for an application of this approach under the assumption of a symmetric null distribution.

More recently, Wang and Ramdas (2022) have shown that e-values (random variables e_i satisfying $E[e_i] \leq 1$ for $i \in \mathcal{H}_0$) can be used as an input to FDR controlling procedures, thereby providing another approach to controlling FDR without the use of classical significance level controlling tests. A notion of average error control for e-values similar to the one used for p-values in the present paper has arisen independently in this literature; see Ren and Barber (2024), Li and Zhang (2025) and Ignatiadis et al. (2025). Interestingly, Ren and Barber (2024) use this idea to draw a connection between e-values and the knockoff literature cited above.

While we are not aware of previous results applying the average significance control criterion (2) to FDR control, the idea of requiring coverage or size control only on average is suggested by empirical Bayes interpretations of the FDR (e.g. Storey, 2002) and anticipated in some discussions in this literature (e.g. Efron, 2007). Subsequent to the first draft of this paper, further results and applications involving FDR control using the average significance level control criterion (2) have been developed by Ignatiadis and Sen (2025), Ignatiadis et al. (2025) and Barber and Samworth (2025).

The rest of this paper is organized as follows. Section 2 introduces the setup and provides an overview of results. Section 3 presents finite sample results and their proofs. Section 4 presents results that are asymptotic in the number m of hypotheses being tested, while the proofs of the asymptotic results are contained in the Supplementary Materials.

2 Setup and Overview of Results

A multiple testing procedure is a function that maps the p-values $p_1, \ldots, p_m$ to a subset $\mathcal{R} = \mathcal{R}(p_1, \ldots, p_m) \subseteq \{1, \ldots, m\}$ of rejected null hypotheses. The false discovery proportion (FDP) of a procedure $\mathcal{R}$ is:

$$\mathrm{FDP}(\mathcal{R}, \mathcal{H}_0) = \frac{\#\left(\mathcal{R} \cap \mathcal{H}_0\right)}{\#\mathcal{R} \vee 1} \tag{4}$$

where $\#\mathcal{A}$ is the cardinality of $\mathcal{A}$ and $a \vee b$ denotes the maximum of a and b. The false discovery rate (FDR) of this procedure is the expectation of the FDP:

$$\mathrm{FDR}(\mathcal{R}, \mathcal{H}_0, P) = E_P \,\mathrm{FDP}(\mathcal{R}, \mathcal{H}_0) = E_P\left[\frac{\#\left(\mathcal{R} \cap \mathcal{H}_0\right)}{\#\mathcal{R} \vee 1}\right] \tag{5}$$

where E_P denotes expectation under the distribution P of the p-values. We say that $\mathcal{R}$ controls the false discovery rate at level q if $\mathrm{FDR}(\mathcal{R}, \mathcal{H}_0, P) \leq q$.

While some of our results are more general, our main focus is on the Benjamini and Hochberg (1995, BH) step-up procedure, and generalizations such as those considered by Benjamini and Yekutieli (2001), Storey (2002) and Blanchard and Roquain (2008). To describe these procedures, let

$$\mathcal{R}_t^{\mathrm{fixed}}(p_1, \ldots, p_n) = \{i : p_i \leq t\}. \tag{6}$$

denote the fixed rejection region procedure with cutoff t. That is, we reject all hypotheses with p-value less than t. Let

$$V(t) = \sum_{i \in \mathcal{H}_0} I(p_i \leq t) = \#\left(\mathcal{R}_t^{\mathrm{fixed}} \cap \mathcal{H}_0\right), \quad S(t) = \sum_{i \notin \mathcal{H}_0} I(p_i \leq t) = \#\left(\mathcal{R}_t^{\mathrm{fixed}} \backslash \mathcal{H}_0\right)$$
$$\text{and } R(t) = V(t) + S(t) = \#\mathcal{R}_t^{\mathrm{fixed}}. \tag{7}$$

The FDP of $\mathcal{R}_t^{\mathrm{fixed}}$ is given by $V(t)/[R(t) \vee 1]$. The BH procedure can be motivated by noting

that, while $V(t)$ cannot be observed, one can form a conservative estimate by replacing it with $m \cdot t$. This gives an estimate of the fixed rejection region FDR:

$$\widehat{\mathrm{FDR}}(t) = \frac{m \cdot t}{\#\mathcal{R}_t^{\mathrm{fixed}} \vee 1} = \frac{m \cdot t}{R(t) \vee 1}. \tag{8}$$

The BH procedure at nominal FDR level q uses a cutoff $\hat{t}_{\mathrm{BH},q}$ based on this estimate:

$$\mathcal{R}_{\mathrm{BH},q}(p_1, \ldots, p_m) = \{i : p_i \leq \hat{t}_{\mathrm{BH},q}\} \quad \text{where} \quad \hat{t}_{\mathrm{BH},q} = \max\{t : \widehat{\mathrm{FDR}}(t) \leq q\}. \tag{9}$$

A more general class of step-up procedures can be formed by using an estimate of the form $\pi m t$ for $V(t)$ and modifying the denominator using a nondecreasing function β, called a shape function:

$$\mathcal{R}_{\pi,\beta(\cdot),q}(p_1, \ldots, p_m) = \{i : p_i \leq \hat{t}_{\pi,\beta(\cdot),q}\} \text{ where } \hat{t}_{\pi,\beta(\cdot),q} = \max\left\{t : \frac{\pi m t}{\beta(R(t))} \leq q\right\}. \tag{10}$$

Such procedures have been considered by, among others, Benjamini and Yekutieli (2001), Storey (2002) and Blanchard and Roquain (2008).

When the p-values satisfy the classical significance level control condition (1), these procedures are known to have the following properties.

(i) The BH procedure controls FDR when p-values are independent (Benjamini and Hochberg, 1995).

(ii) The estimate $\widehat{\mathrm{FDR}}(t)$ is upwardly biased for the FDR of the fixed rejection region procedure $\mathcal{R}_t^{\mathrm{fixed}}$ when p-values are independent (Storey et al., 2004; Liang and Nettleton, 2012).

(iii) The procedure $\mathcal{R}_{1,\beta(\cdot),q}$ (with $\pi = 1$) controls FDR under arbitrary dependence for the shape function $\beta(k) = k\left(\sum_{i=1}^m i^{-1}\right)^{-1}$ (Benjamini and Yekutieli, 2001) and, more generally, when $\beta(k) = \int_0^k x d\nu(x)$ for an arbitrary probability distribution ν on $(0, \infty)$ (Blanchard and Roquain, 2008).

(iv) The BH procedure controls FDR asymptotically (as $m \to \infty$) when the p-values satisfy a weak dependence condition (Storey et al., 2004; Genovese and Wasserman, 2004).

(v) The procedure $\mathcal{R}_{\hat{\pi},\beta(\cdot),q}$, where $\beta(t) = t$ and $\hat{\pi} = (\sum_{i=1}^m I(p_i > \lambda) + 1)/((1-\lambda)m)$ is an estimate of $\#\mathcal{H}_0/m$, controls FDR (a) under fixed m with independent p values using

a slight modification of the procedure (Storey et al., 2004)[1] and (b) asymptotically as $m \to \infty$ when the p-values satisfy a weak dependence condition (Storey et al., 2004; Genovese and Wasserman, 2004).

Our results can be summarized as showing that, when the p-values only satisfy the weaker average significance level control condition (2), properties (ii), (iii) and (iv) continue to hold, but that properties (i) and (v)(a) and (v)(b) in general do not. Section 3.1 shows property (iii) and provides a counterexample to property (i). Section 3.2 shows property (ii). Section 4 shows property (iv). A counterexample for property (v) is given in the Supplementary Materials.

3 Finite Sample Results

This section considers finite sample control of FDR for step-up procedures (Section 3.1) and point estimation of FDR of the fixed rejection region procedure $\mathcal{R}_t^{\text{fixed}}$ (Section 3.2).

3.1 FDR Control

Our result on FDR control for step-up procedures is a corollary of a more general result that uses an invariance assumption on an oracle version of a multiple testing procedure. The basic idea is that, if the p-values satisfy the average significance level control condition (2), then one can form another multiple testing problem in which the classical condition (1) holds by randomly permuting the p-values of the true null hypotheses and multiplying them by $m/\#\mathcal{H}_0$. One can then apply results from the literature to this new setting.

To state our result, we explicitly introduce notation $\mathcal{R}(p_1, \ldots, p_m; \mathcal{H}_0)$ for oracle procedures that depend on the set of true null hypotheses $\mathcal{H}_0$ (typically through the cardinality $\#\mathcal{H}_0$ of this set). We use a permutation invariance condition

$$i \in \mathcal{R}(p_1, \ldots, p_m) \quad \text{iff.} \quad \sigma(i) \in \mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)}) \tag{11}$$

for any permutation σ of the indices $1, \ldots, m$ of the tests. This includes the class of step-up procedures (10), so long as π is either a fixed number or a permutation invariant function of the p-values.

[1]For results with fixed m, Storey et al. (2004) consider a modification in which $\hat{t}_{\pi,\beta(\cdot),q}$ is replaced by λ if $\hat{t}_{\pi,\beta(\cdot),q} > \lambda$.

Theorem 3.1. *Let $\mathcal{R}$ be a multiple testing procedure that satisfies the permutation invariance condition (11), and suppose that the oracle procedure $\widetilde{\mathcal{R}}(p_1, \ldots, p_m; \mathcal{H}_0) = \mathcal{R}(p_1(m_0/m), \ldots, p_m(m_0/m))$ (where $m_0 = \#\mathcal{H}_0$) controls FDR at level q for any $(P, \mathcal{H}_0)$ satisfying the classical significance level control condition (1), regardless of the dependence structure of $p_1, \ldots, p_m$ under P. Then $\mathcal{R}$ controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (2) holds.*

Proof. Given $(P, \mathcal{H}_0)$ such that (2) holds and $p_1, \ldots, p_n$ drawn from P, define $\tilde{p}_i$ as follows. Let σ be a permutation of $\mathcal{H}_0$, taken at random from the set of all permutations of $\mathcal{H}_0$ with equal probability, independently of $p_1, \ldots, p_m$. Extend σ to a permutation on $\{1, \ldots, m\}$ by taking $\sigma(i) = i$ for $i \notin \mathcal{H}_0$. Let $\tilde{p}_i = (m/m_0)p_{\sigma(i)}$, where $m_0 = \#\mathcal{H}_0$. Then, for $i \in \mathcal{H}_0$ and $t \in [0, 1]$,

$$P(\tilde{p}_i \leq t) = \sum_{j \in \mathcal{H}_0} P(\sigma(i) = j) P(p_j(m/m_0) \leq t | \sigma(i) = j) = \frac{1}{m_0} \sum_{j \in \mathcal{H}_0} P(p_j(m/m_0) \leq t)$$

where we use independence of σ and p_j and the fact that $P(\sigma(i) = j) = 1/m_0$. Since $p_1, \ldots, p_m$ satisfy (2) under $(P, \mathcal{H}_0)$, this is bounded by $(m/m_0) \cdot t m_0/m = t$. Thus, letting $\tilde{P}$ denote the distribution of $\tilde{p}_1, \ldots, \tilde{p}_n$ under P, $(\tilde{P}, \mathcal{H}_0)$ satisfies the classical significance level control condition (1). It follows by the assumptions of the theorem that the oracle procedure $\widetilde{\mathcal{R}}(\tilde{p}_1, \ldots, \tilde{p}_m; \mathcal{H}_0) = \mathcal{R}(\tilde{p}_1(m_0/m), \ldots, \tilde{p}_m(m_0/m)) = \mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)})$ controls FDR at level q under $\mathcal{H}_0$ when $p_1, \ldots, p_m$ are drawn according to P. But by permutation invariance of $\mathcal{R}$ and the fact that σ maps $\mathcal{H}_0$ to itself, we have $\#\left(\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)}) \cap \mathcal{H}_0\right) = \#\left(\mathcal{R}(p_1, \ldots, p_m) \cap \mathcal{H}_0\right)$. Also, $\#\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)}) = \#\mathcal{R}(p_1, \ldots, p_m)$ by permutation invariance. Thus, the FDR of $\mathcal{R}(p_m, \ldots, p_m)$ is the same as the FDR of $\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)})$, and is therefore bounded by q. □

As a special case, applying Proposition 2.7 and Lemma 3.2(iii) in Blanchard and Roquain (2008) gives the following.

Corollary 3.1. *The class of dependence controlling step-up procedures of Blanchard and Roquain (2008), given by (10) with $\pi = 1$ and $\beta(r) = \int_0^r x d\nu(x)$ for some probability measure ν, controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (2) holds. In particular, the step-up procedure of Benjamini and Yekutieli (2001), which is given by (10) with $\pi = 1$ and $\beta(r) = r/\left(\sum_{i=1}^m 1/i\right)$, controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (2) holds.*

Key requirements here are that the original procedure (a) controls FDR under arbitrary dependence and (b) can incorporate the m/m_0 adjustment through an oracle result. In particular, (b) rules out procedures of the form $\mathcal{R}_{\hat{\pi},\beta(\cdot),q}$ with $\hat{\pi}$ an estimate of m_0/m, as in Storey (2002). Clearly, ruling out estimates of m_0/m is necessary, since such estimates attempt to use a bound $m_0 \cdot t$ on $V(t)$, whereas average coverage only gives a bound of $m \cdot t$ on the expectation of $V(t)$ (see the Supplementary Materials for a counterexample). Regarding (a), note that even if the p-values satisfy some dependence structure that would guarantee FDR control under classical significance level control (e.g. independence or the positive regression dependency on a subset condition used by Benjamini and Yekutieli (2001)), FDR control is not guaranteed. The following counterexample shows that (a) is necessary in general. In particular, the BH procedure need not control FDR under the average significance level control condition (2) and independent p-values.

Suppose $m \geq 2$ and $q < 2/3$. Let $P(p_1 \leq t) = t \cdot m$ for $0 \leq t \leq (3/2) \cdot (q/m)$ and let $P(p_1 \in ((3/2) \cdot (q/m), 2q/m]) = 0$. Let $P(p_2 \in [a, b]) = (b - a) \cdot m$ for any $(3/2) \cdot (q/m) \leq a \leq b \leq 2q/m$ and let $P(p_2 \in [0, (3/2) \cdot (q/m))) = 0$. We can then distribute the remaining probability mass of p_1, p_2 and $p_3, \ldots, p_m$ over the set $(2q/m, 1]$ so that $\frac{1}{m} \sum_{i=1}^{m} P(p_i \leq t) \leq t$ for all $t \in [0, 1]$ (for example, we can set $p_3, \ldots, p_m$ to be equal to 1 with probability one, and we can set the remaining probability mass for p_1 and p_2 to point masses at 1). Thus, the average significance level condition (2) holds, with $\mathcal{H}_0 = \{1, \ldots, m\}$. Now consider the FDR of the Benjamini-Hochberg procedure, which rejects all hypotheses i such that $p_i \leq q\hat{r}/m$ where $\hat{r}$ is the number of rejected hypotheses. The FDR is equal to the probability of at least one rejection in this case (since $\mathcal{H}_0 = \{1, \ldots, m\}$). Note that the event $p_1 \leq q/m$ implies that hypothesis 1 is rejected, and this has probability q. But the event $q/m < p_1 \leq (3/2) \cdot (q/m)$ and $p_2 \leq 2q/m$ has probability $(q/2) \cdot (q/2)$, and it is disjoint with the event $p_1 \leq q/m$. This gives a lower bound of $q + [(q/m) \cdot (1/2)]^2 > q$ for the FDR. Thus, the FDR is not controlled at level q.

It is worth mentioning here some further results that have been obtained subsequent to the first draft of this paper. First, Ignatiadis et al. (2025) have developed results for e-values using a related notion of average error control: they show that if the e-values $e_1, \ldots, e_m$ satisfy $(1/m) \sum_{i \in \mathcal{H}_0} E[e_i] \leq 1$, then the e-BH procedure of Wang and Ramdas (2022) controls FDR at the nominal level. Using this result and a certain calibration of p-values to e-values, they provide an alternative proof of the main conclusion of Corollary 3.1. Second, Barber and Samworth (2025) have obtained several results characterizing the FDR properties of the BH procedure under (2) and various dependence assumptions on the p-values. In particular,

they show that, when p-values are independent, applying the procedure (10) with $\pi = 1$ and $\beta(r) = r/1.93$ (i.e. the BH procedure at nominal level $q/1.93$) controls FDR. This provides a much less conservative procedure compared to those in Corollary 3.1 in the case of independent p-values.

3.2 Estimation of FDR for Fixed Rejection Region

We now consider using the BH cutoff as an estimate of the FDR for a fixed rejection region multiple testing procedure. Under independent p-values, it is known that $\widehat{\text{FDR}}(t)$ is an upwardly biased estimate of $\text{FDR}(\mathcal{R}_t^{\text{fixed}})$ under the classical significance level control condition (1) (see Storey et al. (2004) and the correction by Liang and Nettleton (2012)). Indeed, a slightly weaker assumption of independence between null and non-null p-values suffices.[2] We now show that this property continues to hold under the weaker average significance level control condition (2). The result essentially follows from the same arguments as in the case where the p-values satisfy the classical significance level control condition.

Theorem 3.2. *Suppose that $(P, \mathcal{H}_0)$ satisfies the average significance level control condition (2) and that the null p-values $\{p_i\}_{i\in\mathcal{H}_0}$ are statistically independent of the non-null p-values $\{p_i\}_{i\notin\mathcal{H}_0}$. Then $E_P\widehat{\text{FDR}}(t) \geq \text{FDR}(\mathcal{R}_t^{\text{fixed}}, \mathcal{H}_0, P)$.*

Proof. For $V(t)$ and $S(t)$ defined in (7), we have

$$E_P\widehat{\text{FDR}}(t) - \text{FDR}(\mathcal{R}_t^{\text{fixed}}, \mathcal{H}_0, P) = E_P\frac{m\cdot t - V(t)}{[V(t)+S(t)]\vee 1} \geq E_P\frac{m\cdot t - V(t)}{[m\cdot t + S(t)]\vee 1}$$

(the last step follows by noting that replacing $V(t)$ with $m\cdot t$ in the denominator weakly decreases the denominator when the numerator is negative and weakly increases the denominator when the numerator is positive). The result then follows by noting that $S(t)$ and $V(t)$ are independent by the independence assumption on p-values, and that $E_PV(t) \leq m\cdot t$ by the assumption that the p-values satisfy the average significance level control condition (2). □

4 Asymptotic Results

We now consider asymptotic FDR control, under a sequence $P = P^{(m)}$ of probability measures and $\mathcal{H}_0 = \mathcal{H}_0^{(m)}$ and $m \to \infty$. We suppress the dependence on m whenever it doesn't

[2] I thank an anonymous referee for pointing this out.

cause confusion, but we note that the p-values form a triangular array, since the distribution (and the set $\mathcal{H}_0$ of true null hypotheses) can change with m. Recall the definitions of $V(t)$, $S(t)$ and $R(t)$ in (7). If the average significance level control condition (2) holds, and the p-values do not exhibit too much statistical dependence, we will have

$$\frac{1}{m}V(t) \leq t + o_P(1) \text{ for all } t \in [0,1]. \tag{12}$$

For some results, we also assume a law of large numbers for the total rejections and rejected true nulls:

$$\frac{1}{m}V(t) \xrightarrow{p} G(t) \leq t \quad \text{and} \quad \frac{1}{m}R(t) \xrightarrow{p} F(t) \text{ for all } t \in [0,1]. \tag{13}$$

These assumptions are analogous to assumptions made for asymptotic FDR control under classical significance level control in the literature (e.g. Storey et al., 2004, Eq. (7)-(9)). The difference here is that the conditions are weaker, since the upper bound in (12) is given by t rather than $t\pi_0$ where π_0 is the limit of $\#\mathcal{H}_0/m$. As one might expect, this will lead to problems for "adaptive" procedures that attempt to estimate π_0 (see the Supplementary Materials for a counterexample). However, as we now show, it is not a problem for the Benjamini-Hochberg procedure, which uses the conservative upper bound of 1. We first show conservative consistency of the BH cutoff (8) for the FDR (and FDP) of the fixed rejection region procedure.

Theorem 4.1. *Let $\widehat{\mathrm{FDR}}(t)$ be the BH estimate, given in (8), of the FDR of the fixed rejection region procedure $\mathcal{R}_t^{fixed}$ given in (6) and suppose that (12) holds. Let $\underline{t}$ be such that there exists $\eta > 0$ with $\frac{1}{m}\sum_{i=1}^m I(p_i \leq \underline{t}) \geq \eta + o_P(1)$. Then*

$$\inf_{t \in [\underline{t},1]} \left[\widehat{\mathrm{FDR}}(t) - \mathrm{FDP}(\mathcal{R}_t^{fixed}, \mathcal{H}_0)\right] \geq o_P(1).$$

If, in addition, (13) holds for continuous functions G and F, then, letting $\mathrm{FDR}_\infty(t) = G(t)/F(t)$, we have

$$\sup_{t \in [\underline{t},1]} \left|\mathrm{FDP}(\mathcal{R}_t^{fixed}, \mathcal{H}_0) - \mathrm{FDR}_\infty(t)\right| \xrightarrow{p} 0, \quad \sup_{t \in [\underline{t},1]} \left|\mathrm{FDR}(\mathcal{R}_t^{fixed}, \mathcal{H}_0, P) - \mathrm{FDR}_\infty(t)\right| \to 0$$
$$\textit{and } \inf_{t \in [\underline{t},1]} \left[\widehat{\mathrm{FDR}}(t) - \mathrm{FDR}(\mathcal{R}_t^{fixed}, \mathcal{H}_0, P)\right] \geq o_P(1).$$

The proof of Theorem 4.1 is given in the Supplementary Materials. Next, we state a

result showing asymptotic control of FDR for the BH procedure $\mathcal{R}_{\mathrm{BH},q}$ defined in (9).

Theorem 4.2. *Suppose Assumptions (12) and (13) hold for continuous functions G and F and that there exists $t^* > 0$ such that $F(t^*) > 0$ and $G(t^*)/F(t^*) < q$. Then*

$$\mathrm{FDP}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0) \leq q + o_P(1) \quad \textit{and} \quad \mathrm{FDR}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0, P) \leq q + o(1).$$

The proof of Theorem 4.2 is given in the Supplementary Materials. The condition on t^* used in Theorem 4.2 imposes a lower bound on the proportion of total rejections relative to null rejections at nominal level t^*. This condition requires that the hypothesis tests have sufficient power on average, and that the proportion $(m - \#\mathcal{H}_0)/m$ of non-null hypotheses is not too small as $m \to \infty$. Similar conditions have been used in the asymptotic analysis of multiple hypothesis testing procedures under classical significance level control (e.g. Storey et al., 2004, Theorem 4).

References

Arias-Castro, E. and S. Chen (2017): "Distribution-free multiple testing," *Electronic Journal of Statistics*, 11, 1983–2001.

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2022): "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 90, 2567–2602.

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

Barber, R. F. and R. J. Samworth (2025): "False discovery rate control with compound p-values," ArXiv:2507.21465 [math].

Benjamini, Y. and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Benjamini, Y. and D. Yekutieli (2001): "The Control of the False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, 29, 1165–1188.

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

Cai, T. T., M. Low, and Z. Ma (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109, 1054–1070.

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

Efron, B. (2007): "Size, power and false discovery rates," *The Annals of Statistics*, 35, 1351–1377.

Genovese, C. and L. Wasserman (2004): "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061.

Ignatiadis, N. and B. Sen (2025): "Empirical partially Bayes multiple testing and compound χ^2 decisions," *The Annals of Statistics*, 53, 1–36.

Ignatiadis, N., R. Wang, and A. Ramdas (2025): "Asymptotic and compound e-values: multiple testing and empirical Bayes," .

Li, G. and X. Zhang (2025): "A note on e-values and multiple testing," *Biometrika*, 112, asae050.

Liang, K. and D. Nettleton (2012): "Adaptive and dynamic adaptive procedures for false discovery rate control and estimation," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74, 163–182.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25, 2547–2554.

Nychka, D. (1988): "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143.

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

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

Storey, J. D., J. E. Taylor, and D. Siegmund (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, 187–205.

Wahba, G. (1983): "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society: Series B (Methodological)*, 45, 133–150.

Wang, R. and A. Ramdas (2022): "False Discovery Rate Control with E-values," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 822–852.

Wasserman, L. (2007): *All of Nonparametric Statistics*, New York: Springer.