A unified framework for bandit multiple testing

Abstract

In bandit multiple hypothesis testing, each arm corresponds to a different null hypothesis that we wish to test, and the goal is to design adaptive algorithms that correctly identify large set of interesting arms (true discoveries), while only mistakenly identifying a few uninteresting ones (false discoveries). One common metric in non-bandit multiple testing is the false discovery rate (FDR). We propose a unified, modular framework for bandit FDR control that emphasizes the decoupling of exploration and summarization of evidence. We utilize the powerful martingale-based concept of "e-processes" to ensure FDR control for arbitrary composite nulls, exploration rules and stopping times in generic problem settings. In particular, valid FDR control holds even if the reward distributions of the arms could be dependent, multiple arms may be queried simultaneously, and multiple (cooperating or competing) agents may be querying arms, covering combinatorial semi-bandit type settings as well. Prior work has considered in great detail the setting where each arm's reward distribution is independent and sub-Gaussian, and a single arm is queried at each step. Our framework recovers matching sample complexity guarantees in this special case, and performs comparably or better in practice. For other settings, sample complexities will depend on the finer details of the problem (composite nulls being tested, exploration algorithm, data dependence structure, stopping rule) and we do not explore these; our contribution is to show that the FDR guarantee is clean and entirely agnostic to these details.

1 Introduction to bandit multiple hypothesis testing

Scientific experimentation is often a sequential process. To test a single null hypothesis — with "null" capturing the setting of no scientific interest, and the alternative being scientifically interesting — scientists often collect an increasing amount of experimental data in order to gather sufficient evidence such that they can potentially reject the null hypothesis (i.e. make a scientific discovery) with a high degree of statistical confidence. As long as the collected evidence remains thin, they do not reject the null hypothesis and do not proclaim a discovery. Since executing each additional unit of data (stemming from an experiment or trial) has an associated cost (in the form of time, money, resources), the scientist would like to stop as soon as possible. This becomes increasingly prevalent when the scientist is testing multiple hypotheses at the same time, and investing resources into testing one means divesting it from another.

For example, consider the case of a scientist at a pharmaceutical company who wants to discover which of several drug candidates under consideration are truly effective (i.e. testing a hypothesis of whether each candidate has greater than baseline effect) through an adaptive sequential assignment of candidates to participants. Performing follow up studies on each discovery is expensive, so the scientist does not want to make many "false discoveries" i.e. drugs that did not have an actual effect, but were proclaimed to have one by the scientist. To achieve these goals, one could imagine the scientist collecting more data for candidates whose efficacy is unclear but appear promising (e.g. drugs with nontrivial but inconclusive evidence), and stop sampling candidates that have relatively clear results already (e.g. drugs that have a clear and large effect, or seemingly no effect).

Past work. This problem combines the challenges of multiple hypothesis testing with multi-arm bandits (MABs). In a "doubly-sequential" version of the problem studied by Yang et al. (2017), one

Preprint. Under review.

encounters a sequence of MAB problems over time. Each MAB was used to test a single special placebo arm against several treatment arms, and if at least one treatment dominated the placebo, then they aimed to return the best treatment. Thus each MAB was itself a single adaptive sequential hypothesis test, and the authors aimed not to make too many false discoveries over the sequence of MAB instances.

This paper instead considers the formulation of Jamieson and Jain (2018), henceforth called JJ, but our techniques apply equally well to the above setup. *(Note: our setup is very different from the active classification work of Jain and Jamieson (2019).)* To recap, JJ consider a single MAB instance without a placebo arm (or rather, leaving it implicit), and try to identify as many treatments that work better than chance as possible, without too many false identifications. To clarify, we associate each arm with one null hypothesis, for example being that the corresponding drug has no (significant) effect. A single observed reward when pulling an arm corresponds to a statistic that summarizes the results of one experiment with the corresponding drug, and the average reward across many experiments could correspond to an estimate of the average treatment effect (which would be zero for null arms and positive for non-nulls). Thus, a strategy for for quickly finding the arms with positive means corresponds to a strategy for allocating trial patients to drug candidates that allows the scientists to rapidly find the effective drugs.

However, the above corresponds to only the simplest problem setting. In more complex settings, it may be possible to pull multiple arms in each round, and observe correlated rewards. Further, the arms may have some combinatorial structure that allows only certain subsets of arms to be pulled. There could be multiple agents (eg: hospitals) pulling the same set of arms and seeing independent rewards (eg: different patients) or dependent rewards (eg: patient overlap or interference). Further, if some set of experiments by one scientist yielded suggestive but inconclusive evidence, another may want to follow up, but not start from scratch, instead picking up from where the first left off. Last, the MAB may be stopped for a variety of reasons that may or may not be in the control of the scientist (eg: a faster usage of funding than expected, or additional funding is secured). We dive in the details of these scenarios in Appendix B.

Our contribution. We introduce a modular meta-algorithm for bandit multiple testing that utilizes "e-values" — or, more appropriately, their sequential analog, e-processes — a recently introduced alternative to p-values (or p-processes) for various testing problems, and inherently related to martingales, gambling and betting (Vovk and Wang, 2020; Shafer, 2020; Grünwald et al., 2020; Howard et al., 2020; Wang and Ramdas, 2020). Utilizing e-processes provide our meta-algorithm with several benefits. (a) For composite nulls, it is typically easier to construct e-processes than p-processes; the same holds when data from a single source is dependent. When combining evidence from disparate (independent or dependent) sources, it is also more straightforward to combine e-values than p-values (see Appendix B). (b) The same multiple testing step applies in all bandit multiple testing problems, regardless of all the various details of the problem setup mentioned in the previous paragraph. This is not true when working for p-values. (c) The exploration step can be — but does not have to be — decoupled from the multiple testing (combining evidence) step. This results in a modular procedure that can be easily ported to new problem settings to yield transparent guarantees on FDR control.

By virtue of being a meta-algorithm, we do not (and cannot) provide "generic" sample complexity guarantees: these will depend on all of the finer problem details mentioned above, on the exploration algorithm employed, on which e-processes are constructed. Our emphasis is on the flexibility with which FDR control can be guaranteed in a vast variety of problem setups. Further research can pick up one problem at a time and design sensible exploration strategies and stopping rules, developing sampling complexity bounds for each, and these bounds will be inherited by the meta-algorithm. However, we do formulate some generic exploration algorithms in Appendix B based on best arm identification algorithms (Audibert and Bubeck, 2010; Kalyanakrishnan et al., 2012; Chen et al., 2014; Jamieson et al., 2014; Kaufmann et al., 2016; Chen et al., 2017; Jourdan et al., 2021).

When instantiated to the particular problem setup studied by JJ (independent, sub-Gaussian rewards, one arm in each round, etc.), we get a slightly different algorithm from them — the exploration strategy can be inherited to stay the same, but the multiple testing part differs. JJ use p-processes for each arm to determine whether that arm should be added to the rejection set, and correct for testing multiple hypotheses by using the BH procedure (Benjamini and Hochberg, 1995) to ensure that the **false discovery rate (FDR)**, i.e. the proportion of rejections that are false discoveries in expectation, is controlled at some fixed level δ. Adaptive sampling induces a peculiar form of dependence amongst the p-values, for which the BH procedure provides error control at an inflated level; in other words, one has to use BH at a more stringent level of approximately $\delta / \log(16/\delta)$ to ensure that the FDR is less than δ. On the other hand, we use the e-BH procedure (Wang and Ramdas, 2020), an analogous procedure for e-values, which can ensure the FDR is less than δ without any inflation, regardless of the dependence structure between the e-values of each arm. Our algorithm has improved sample efficiency in simulations and the same sample complexity in theory.

Formal bandit setting. We define the bandit as having k arms, and ν_i as the (unknown) reward distribution for arm $i \in [k] = \{1,...,k\}$. Every arm i is associated with a null hypothesis, which is represented by a known, prespecified set of distributions $\mathcal{P}_i$. If $|\mathcal{P}_i| = 1$, it is a 'point null hypothesis', and otherwise it is a 'composite null hypothesis'. Examples of the latter include "all $[0,1]$-bounded distributions with mean ≤ 0.5" or "all 1-sub-Gaussian distributions with mean ≤ 0" or "all distributions that are symmetric around 0" or "all distributions with median ≤ 0". If $\nu_i \in \mathcal{P}_i$, then we say that the i-th null hypothesis is true and we call i a null arm; else, we say i-th null hypothesis is false and we call it a non-null arm. Thus, the set of arms are partitioned into two disjoint sets: nulls $\mathcal{H}_0 \subseteq [k]$ and non-nulls $\mathcal{H}_1 \equiv [k] \setminus \mathcal{H}_1$.

Let $\mathcal{K} \subseteq 2^{[k]}$ denote the subsets of arms that can be jointly queried in each round. At each time t, the algorithm chooses a subset of arms $\mathcal{I}_t \subset [k]$ such that $\mathcal{I}_t \in \mathcal{K}$ to sample jointly from. The special choice of $\mathcal{K} = \{\{1\},\{2\},...,\{k\}\}$ recovers the standard bandit setup, but otherwise this setting is known as combinatorial bandits with semi-bandit feedback (Chen et al., 2016). We denote the reward sampled at time t from arm $i \in \mathcal{I}_t$ as $X_{i,t}$. Let $T_i(t)$ denote the number of times arm i has been sampled by time t, and $t_i(j)$ be the time of the jth sample from arm i.

We now define a canonical "filtration" for our bandit problem. A filtration $(\mathcal{F}_t)_{t \geq 0}$ is a series of nested sigma-algebras that encapsulates what information is known at time t. (We drop the subscript and just write $(\mathcal{F}_t)$ for brevity, and drop the parentheses when just referring to a single sigma-algebra at time t.) Define the **canonical filtration** as follows for $t \in \mathbb{N}$: $\mathcal{F}_t \equiv \sigma(U \cup \{(i,j,X_{i,j}) : j \leq t, i \in \mathcal{I}_j\})$ and we let $\mathcal{F}_0 \equiv \sigma(U)$ where U is uniformly distributed on $[0,1]$ and its bits capture all private randomness used by the bandit algorithm that are independent of all observed rewards. Let (λ_t) be a sequence of random variables indexed by $t \in \mathbb{N}$. (λ_t) is said to be **predictable** w.r.t. $(\mathcal{F}_t)$ if λ_t is measurable w.r.t. $\mathcal{F}_{t-1}$ i.e. λ_t is fully specified given the information in $\mathcal{F}_{t-1}$. An $\mathbb{N}$-valued random variable τ is a **stopping time** (or stopping rule) w.r.t. to $(\mathcal{F}_t)$ if $\{\tau = t\} \in \mathcal{F}_t$ — in other words, at each time t, we know whether or not stop collecting data. Let $\mathcal{T}$ denote the set of all possible stopping times/rules w.r.t. $(\mathcal{F}_t)$, potentially infinite.

Technically, the algorithm must not just specify a strategy to select $\mathcal{I}_t$, but also specify when sampling will stop. This is denoted by the stopping rule or stopping time $\tau^* \in \mathcal{T}$.

Once the algorithm halts at some time τ, it produces a rejection set $\mathcal{S}_\tau \subseteq [k]$. We consider two metrics w.r.t. $\mathcal{S}$: the FDR as discussed prior, and **true positive rate** (TPR), which is the proportion of non-nulls that are discovered in expectation, which we define as follows:

$$\mathrm{FDR}(\mathcal{S}_\tau) \equiv \mathbb{E}\left[\frac{|\mathcal{H}_0 \cap \mathcal{S}_\tau|}{|\mathcal{S}_\tau| \vee 1}\right], \qquad \mathrm{TPR}(\mathcal{S}_\tau) \equiv \mathbb{E}\left[\frac{|\mathcal{H}_1 \cap \mathcal{S}_\tau|}{|\mathcal{H}_1|}\right].$$

We consider algorithms that always satisfy $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ for any number and configuration of nulls $\mathcal{H}_0$ and any choice of null and non-null distributions. In fact, our algorithm will produce a sequence of candidate rejection sets $(\mathcal{S}_t)$ that satisfies $\sup_{\tau \in \mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$. This is a much stronger guarantee than the typical setting considered in the multiple testing literature.

In hypothesis testing, the set of null distributions $\mathcal{P}_i$ for each arm i is known, because the user defines the null hypothesis they are interested in testing. *When the null hypothesis is false, the non-null distribution can be arbitrary*. Consequently, we can prove results about FDR, but we cannot prove guarantees about TPR without several further assumptions on the non-null distributions, dependence across arms, etc.

Finally, note that in bandit multiple testing, one does not care about regret. The problem is more akin to *pure exploration*, where we aim to find a $\mathcal{S}$ with $\mathrm{FDR}(\mathcal{S}_{\tau^*}) \leq \delta$ and large TPR as quickly as possible.

2 E-processes versus p-processes

A **e-variable**, E, is a nonnegative random variable where $\mathbb{E}[E] \leq 1$ when the null hypothesis is true. In contrast, the more commonly used **p-variable**, P, is defined to have a support on $(0,1)$ and satisfy $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ when the null hypothesis is true. To clearly delineate when we are discussing solely the properties of a random variable, we also use the terms "e-value" e and "p-value" p to refer the realized values of a e-variable E and a p-variable P (their instantiations on a particular set of data). E-variables and p-variables are connected through Markov's inequality, which implies that $1/E$ is a p-variable (but $1/P$ is not in general an e-variable). Rejecting a null hypothesis is usually based on observing a small p-value or a large e-value. For example, to control the false positive rate at 0.05 for a single hypothesis test, we reject the null when $p \leq 0.05$ or when $e \geq 20$.

Since bandit algorithms operate over time, we define sequential versions of p-variables and e-variables. A **p-process**, denoted $(P_t)_{t\geq 1}$, is a sequence of random variables such that $\sup_{\tau\in\mathcal{T}}\mathbb{P}(P_\tau\leq\alpha)\leq\alpha$ for any $\alpha\in(0,1)$. In contrast, an **e-process** $(E_t)_{t\geq 1}$ must satisfy $\sup_{\tau\in\mathcal{T}}\mathbb{E}[E_\tau]\leq 1$. These sequentially valid forms of p-variables and e-variables are crucial since we allow the bandit algorithm to choose in a data-dependent manner when to stop and output a rejection set. Thus, we must ensure the respective properties of p-variables and e-variables hold over all stopping times.

These concepts are intimately tied to sequential testing and sequential estimation using confidence sequences (Ramdas et al., 2020), but most important for is the central role of nonnegative (super)martingales in the construction of efficient e-processes. To summarize, (a) for point nulls, all admissible e-processes are simply nonnegative martingales, and the safety property follows from the optional stopping theorem, (b) for composite nulls, admissible e-processes are either nonnegative martingales, or nonnegative supermartingales, or the infimum (over the distributions in the null) of nonnegative martingales. Associated connections to betting (Waudby-Smith and Ramdas, 2021) are also important for development of sample efficient algorithms and we discuss how we use betting ideas in Appendix D.

Why use e-processes over p-processes? Wang and Ramdas (2020) describe a multitude of advantages outside of the bandit setting; these advantages also apply to the bandit setting but we do not redescribe them here for brevity. However, we will describe multiple ways in which using e-variables instead of p-variables as a measure of evidence in the bandit setting allows for powerful algorithms in both flexibility of usage and sample complexity. While this question has been the focus of a recent line of work for hypothesis tests in general (Shafer, 2020; Vovk and Wang, 2021; Grünwald et al., 2020; Wang and Ramdas, 2020), we will explore how the properties of e-variables allow us to consider novel bandit setups and algorithms. In particular, e-variables allow us to account for the dependency between statistics computed for each arm without additional correction. Further, we explore how e-processes and p-processes can be merged under different conditions in Appendix B to facilitate incorporation of existing evidence and cooperation between multiple agents and present concrete ways to construct e-processes.

Since any non-trivial bandit algorithm will base its sampling choice on the rewards attained so far for every arm, average rewards of each arm are biased and dependent on each other in complex ways even if the algorithm is stopped at a fixed time Nie et al. (2018); Shin et al. (2021, 2019, 2020). Even under a non-adaptive uniform sampling rule, an adaptive stopping rule can induce complex dependencies between reward statistics of each arm. When using both adaptive sampling and stopping, the dependence effects are only compounded. Nevertheless, using e-variable based algorithms for maintaining a rejection set can allow us to prove FDR guarantees simply without any assumptions on the sampling method. This is contrast with procedures involving p-variables, such as the ones used in JJ, which require the test level to be corrected by a factor of at least $\log(1/\alpha)$ when p-variables are independent in the limit of infinite sampling, and a factor of $\log k$ otherwise. We will explain when and why p-variables require this additional correction in Section 3.

3 Multiple testing procedures with FDR control

We now introduce two multiple testing procedures that output a rejection set with provable FDR control. We will first describe the guarantees provided by the BH procedure (Benjamini and Hochberg, 1995), a classic multiple testing procedure that operates on p-variables. Then, we will describe e-BH, the e-variable analog of BH. Our key message in this section is that classical BH will have looser or tighter control of the FDR based upon the dependence structure of the p-variables it is operating on. On the other hand, e-BH provides a consistent guarantee on the FDR even when the e-variables are arbitrarily dependent. Both procedures take an input parameter $\alpha\in(0,1)$ that controls degree of FDR guarantee.

Benjamini-Hochberg (BH). A set $\mathcal{S}$ of p-values is called **p-compliant** (Su, 2018) at level α if:

$$\max_{i\in\mathcal{S}} p_i\leq\tfrac{|\mathcal{S}|\alpha}{k}. \tag{1}$$

The BH procedure with input $p_1,\ldots,p_k$ outputs the largest p-compliant set w.r.t. the input, which we denote $\mathrm{BH}[\alpha](p_1,\ldots,p_k)$. We must also define a condition on the joint distribution of $P_1,\ldots,P_k$, which is called **positive regression dependence on subset (PRDS)**. We provide a formal definition in Appendix E, and it is sufficient to think of this condition as positive dependence between $P_1,\ldots,P_k$, with independence being a special case. Now, we describe the FDR control of the BH procedure.

Fact 1 (BH FDR control. Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001)**.** *Let* $\mathcal{S}=\mathrm{BH}[\alpha](p_1,\ldots,p_k)$*. If* $P_1,\ldots P_k$ *are PRDS, then* $\mathrm{FDR}(\mathcal{S})\leq\alpha$*. Otherwise, under arbitrary dependence amongst* $P_1,\ldots P_k$*, the BH procedure ensures* $\mathrm{FDR}(\mathcal{S})\leq\alpha\ell_k$*, where* $\ell_k\equiv\sum_{i=1}^k 1/k\approx\log k$.

Thus, in the case of arbitrary dependence, the FDR control of BH is larger by a factor of $\ell_k \approx \log k$.

Fact 2 (P-compliant p-values FDR control. Su 2018)**.** *If $\mathcal{S}$ is p-compliant at level α and $P_1,\dots,P_k$ satisfy PRDS*[1]*, then* $\mathrm{FDR}(\mathcal{S}) \le \alpha(1+\log(1/\alpha))$*. Otherwise, when there is arbitrary dependence among $P_1,\dots,P_k$,* $\mathrm{FDR}(\mathcal{S}) \le \alpha(1+\log(1/\alpha))\ell_k \approx \alpha(1+\log(1/\alpha))\log k$.

These two facts do not imply each other; the BH procedure outputs the largest compliant set and hence gets away with a stronger error guarantee in both cases. We can use the compliance property to prove FDR control when applying the BH procedure to stopped p-values in some bandit problem setups. We elaborate in Section 5 on how Fact 2 is utilized, and in what settings these dependence structures arise.

e-BH has no correction. The e-BH procedure created by Wang and Ramdas (2020) uses e-variables instead of p-variables and proceeds similarly to the BH procedure. In this case, let $e_1,\dots,e_k$ be the realized e-values for a set of e-variables $E_1,\dots,E_k$. Define $e_{[i]}$ to be the ith largest e-value for $i \in [k]$. A set $\mathcal{S}$ is **e-compliant** at level α iff $\mathcal{S}$ satisfies the following:

$$\min_{i \in \mathcal{S}} e_i \ge \tfrac{k}{\alpha|\mathcal{S}|}. \tag{2}$$

The e-BH procedure outputs the largest e-compliant set, which we denote as $\mathrm{eBH}[\alpha](e_1,\dots,e_k)$. For e-variables, the same guarantee applies for all e-compliant sets and under all dependence structures.

Fact 3 (E-variable compliance FDR control. Wang and Ramdas 2020)**.** *If $\mathcal{S}$ is e-compliant at level α, then* $\mathrm{FDR}(\mathcal{S}) \le \alpha$ *regardless of the dependence structure.*

This is a significant advantage for using e-BH and e-variables, since the FDR control does not change under different types of dependence, unlike the extra factors of approximately $\log(1/\alpha), \log k$ and $\log(1/\alpha)\log k$ paid by p-variables in different settings.

In the case where p-variables can only be constructed as $P = 1/E$, where E is e-variable, the rejection sets output by BH and e-BH are identical. However, the e-variable compliance guarantee in Fact 3 provides identical or tighter FDR control than the BH procedure guarantee in Fact 1 or p-variable compliance guarantee in Fact 2. Thus, e-variables and e-BH offer a degree of robustness against arbitrary dependence, since any algorithm using e-BH does not have to adjust α to guarantee the same level of $\mathrm{FDR}(\mathcal{S}) \le \delta$ for a fixed δ under different dependence structures.

4 Decoupling exploration and evidence: a unified framework

We propose a framework for bandit algorithms that separates each algorithm into an **exploration** component and an **evidence** component; Algorithm 1 specifies a meta-algorithm combining the two.

Algorithm 1: A meta-algorithm for bandit multiple testing that decouples exploration and evidence. The evidence component can track p-processes or e-processes for each arm and use BH or e-BH.

```
Input: Exploration component $(\mathcal{A}_t)$, stopping rule $\tau^*$, desired level of FDR control $\delta \in (0,1)$. Let
        $\delta'$ be the correction of $\delta$ for BH based upon the dependencies of $X_{1,t},\dots,X_{k,t}$. Set $D_0 = \emptyset$.
for $t$ in $1\dots$ do
    $\mathcal{I}_t := \mathcal{A}_t(D_{t-1}) \subseteq [k]$
    Obtain rewards for each $i \in \mathcal{I}_t$, and update data $D_t := D_{t-1} \cup \{(i,t,X_{i,t}) : i \in \mathcal{I}_t\}$.
    Update e-process or p-process for each queried arm (summarizing evidence against each null).
    $\mathcal{S}_t := \begin{cases} \mathrm{BH}[\delta'](p_{1,t},\dots,p_{k,t}) \text{ or arbitrary p-compliant set} & \text{if using p-variables} \\ \mathrm{eBH}[\delta](e_{1,t},\dots,e_{k,t}) \text{ or arbitrary e-compliant set} & \text{if using e-variables} \end{cases}$
    if $\tau^* = t$ then stop and return $\mathcal{S}_t$;
end
```

The exploration component is a sequence of functions $(\mathcal{A}_t)$, where $\mathcal{A}_t : \mathcal{F}_{t-1} \mapsto \mathcal{K}$ specifies the queried arms $\mathcal{I}_t := \mathcal{A}_t(D_{t-1})$, and $D_t := \{(i,j,X_{i,j}) : j < t, i \in \mathcal{I}_j\}$ is the observed data. $\mathcal{A}_t$ is "non-adaptive" if it does not depend on the data, but only on some external randomness U. Regardless of how the exploration component $(\mathcal{A}_t)$ is constructed, our framework guarantees that $\mathrm{FDR}(\mathcal{S}) \le \delta$ for a fixed δ. Similarly, τ^* is adaptive if it depends on the data, and is not determined purely by U.

Evidence component. The FDR control provided by Algorithm 1 is solely due to the procedure for constructing the rejection set in the evidence component, which is completely agnostic to $(\mathcal{A}_t)$. To

[1] Su (2018) technically employs a *slightly* weaker condition which implies PRDS.

do so, the evidence component outputs $\mathcal{S}_t \subseteq [k]$ at each time step t such that $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ for any stopping time τ. If we keep track of a p-process $(P_{i,t})$ for each arm $i \in [k]$, then applying the BH procedure at every time step, including at τ, would yield a $\mathcal{S}$ that has FDR control, potentially at some adjusted level. In contrast, we can apply e-BH to our e-processes $(E_{i,t})$ at each time, including the stopping time (since stopped e-processes are e-variables by definition), and inherit the guarantee from Fact 3 which always gives FDR control at the desired level.

Thus, this framework allows us to guarantee $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ for any stopping time $\tau \in \mathcal{T}$, agnostic of the exploration component. We discuss some generic exploration strategies in Appendix B. In the next section, we will formalize these guarantees and discuss the benefits afforded by using e-variables and e-BH in this framework instead of p-variables and BH.

5 Dependence structures

In this section, we discuss the different dependence structures that arise under different bandit settings and the choice of δ' that ensure FDR control at δ in the p-variable and BH case. Table 1 summarizes the guarantees and choices for each type of dependence. Prior work on hypothesis testing in the bandit setting by JJ has only considered the non-combinatorial bandit case where $X_{1,t},\ldots,X_{k,t}$ are independent. Critically, JJ employ BH and p-variables in their algorithm, and the FDR guarantee of BH changes based on these dependence structures. On the other hand, choosing $\alpha = \delta$ for e-BH is sufficient to guarantee FDR control at level δ for any type of dependence between e-variables, but only sufficient for BH in the non-adaptive, PRDS $X_{1,t},\ldots,X_{k,t}$ setting. We show that there is a wide range of dependence structures that require different degrees of correction for BH. Specifically, we will set an appropriate choice of δ' in each of these situations such that Algorithm 1 with p-variables can ensure FDR control level δ. We include proofs of all results in this section in Appendix A.1.

Table 1: FDR control guaranteed by BH, and the δ' to ensure δ control of FDR in Algorithm 1 under different dependence structures and adaptivity of $(\mathcal{A}_t)$. Adaptivity and arbitrary dependence both require extra correction for BH, but e-BH provides $\mathrm{FDR}(\mathcal{S}) \leq \alpha$ in all settings in the table.

	Dependence of $X_{1,t},\ldots,X_{k,t}$	
Adaptivity of $(\mathcal{A}_t)$ and τ^*	*independent*	*arbitrarily dependent*
non-adaptive	$\mathrm{FDR}(\mathcal{S}) \leq \alpha$ $\delta' = \delta$	$\mathrm{FDR}(\mathcal{S}) \leq \alpha \log n$ $\delta' = \delta / \log k$ (Prop. 2)
adaptive	$\mathrm{FDR}(\mathcal{S}) \leq \alpha(1+\log(1/\alpha)) \wedge \alpha \log k$ $\delta' = c_\delta \wedge \delta \log k$ (Prop. 1)	

Adaptive $(\mathcal{A}_t)$ **and independent** $X_{1,t},\ldots,X_{k,t}$**.** JJ consider this case in the non-combinatorial bandit setting, but their insights also techniques also generalize to the combinatorial setting. In the language of compliance (not explicitly used in JJ), JJ make the key insight that running BH on the p-variables for each arm produces a rejection set that is actually p-compliant with a different set of independent p-variables. Define $P_1^*,\ldots,P_k^*$, where $P_i^* = \inf_{t \in \mathbb{N}} P_{i,t}$ for each $i \in [k]$ i.e. each arm's p-variable in the infinite sample limit. Since $(P_{i,t})$ is an p-process for each arm $i \in [k]$, the corresponding P_i^* is a p-variable. Further, $P_1^*,\ldots,P_k^*$ are independent because $X_{1,t},\ldots,X_{k,t}$ are independent. By definition of $P_1^*,\ldots,P_k^*$, $p_{i,t} \leq p_i^*$ for any $i \in [k]$ and any $t \in \mathbb{N}$. Thus, $\mathcal{S}_{\tau^*}$ is p-compliant w.r.t. $p_1^*,\ldots,p_k^*$, and has its FDR bounded under $\alpha(1+\log(1/\alpha))$ by courtesy of Fact 2. At the same time, the arbitrary dependence guarantee from Fact 1 still applies. This gives us the following guarantee:

Proposition 1. *When* $(\mathcal{A}_t)$ *is adaptive and* $X_{1,t},\ldots,X_{k,t}$ *are independent, Algorithm 1 with p-processes and BH guarantees* $\sup_{\tau \in \mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ *if* $\delta' \leq c_\delta \wedge (\delta / \ell_k)$*, where for any* $\delta \in (0,1)$*, define* $c_\delta \leq \delta$ *as the solution to*

$$c_\delta(1+\log(1/c_\delta)) = \delta.$$

JJ prove a similar bound to Proposition 1. However, they used a larger FDR bound for p-compliant sets with worse constants (which was subsequently improved by Su (2018) as presented earlier), and they only considered the non-combinatorial case. Proposition 1 uses an optimal bound on p-compliant sets from Fact 2, and is valid in our combinatorial bandit setup.

Adaptive $(\mathcal{A}_t)$ **and arbitrarily dependent** $X_{1,t},\ldots,X_{k,t}$**.** In the general combinatorial bandit setting, where the algorithm chooses a subset of arms or "superarm" at each time to jointly sample from, we will have multiple samples from multiple arms in the same time step, and

$X_{1,t},\dots,X_{k,t}$ can be arbitrarily dependent. Consequently, the p-variables corresponding to each arm can also be arbitrarily dependent. For example, a superarm could consist of all arms, and the sampling rule could be to simplly sample this superarm that encompasses all arms. Then, the p-variable distribution would directly depend on the reward distribution of the arms. Thus, we can provide the following guarantee by Fact 1 when using p-variables as a result of Fact 3.

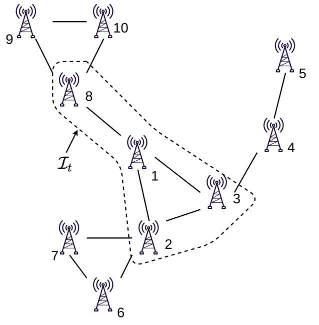


Figure 1: A superarm consists of a node and all its neighbors. The dotted line captures $\mathcal{I}_t$, the superarm around node 1.

Proposition 2. *When $(\mathcal{A}_t)$ is adaptive and $X_{1,t},\dots,X_{k,t}$ are dependent, Algorithm 1 with p-variables and BH guarantees* $\mathrm{FDR}(\mathcal{S}_\tau)\leq\delta$ *for any stopping time $\tau\in\mathcal{T}$ if $\delta'\leq\delta/\log k$.*

Finally, consider a setting structured setting where we cannot output the rejection set of BH. This often occurs in directed acyclic graph settings where there is a hierarchy among hypotheses that constraints the allowed rejecion sets (Ramdas et al., 2019; Lei et al., 2020). Instead, we would like to output the largest compliant set that respects the structural constraints. By Fact 2, we get the following FDR control.

Proposition 3. *If $(\mathcal{A}_t)$ is adaptive and $X_{1,t},\dots,X_{k,t}$ are dependent, Algorithm 1 with p-variables that outputs an arbitrary p-compliant $\mathcal{S}_t$ guarantees* $\mathrm{FDR}(\mathcal{S}_\tau)\leq\delta$ *for any stopping time $\tau\in\mathcal{T}$ if $\delta'\leq c_{\delta/\log k}$.*

We explore the structured setting with greater depth in Appendix B. Unlike p-variables, e-variables do not need correction in any setting.

Proposition 4. *When $(\mathcal{A}_t)$ is adaptive and $X_{1,t},\dots,X_{k,t}$ are dependent, Algorithm 1 with e-variables, which runs e-BH at level δ or outputs a e-compliant set at level δ, guarantees* $\sup_{\tau\in\mathcal{T}}\mathrm{FDR}(\mathcal{S}_\tau)\leq\delta$.

Thus, using e-variables and e-BH (or any e-compliant set) with parameter δ is valid for any choice of $(\mathcal{A}_t)$ and any dependence. We give an example where $X_{1,t},\dots,X_{k,t}$ might be arbitrarily dependent.

Example: sampling nodes on a graph. A scenario where $X_{1,t},\dots,X_{k,t}$ may naturally have dependence is when each arm corresponds to a node on a graph. The superarms in this situation could be defined w.r.t. to a graph constraint e.g. "two nodes connected by an edge" or "a node and its neighbors". Graph bandits has been studied in the regret setting and have many real world applications (Valko, 2016). We could imagine a scenario where low power sensors in a sensor network can only communicate locally. A centralized algorithm is tasked with querying the sensors to find those with high activity. A sensor may only provide activity information about itself and nearby sensors, which can be arbitrarily dependent. Figure 1 illustrates a superarm in this situation. Thus, if Proposition 2 is used to guarantee $\mathrm{FDR}(\mathcal{S})\leq\delta$ with p-variables, it pays a $\log k$ correction, while Proposition 4 can guarantee e-variables need no correction. We simulate this setting in Appendix C, and show these differences empirically.

Other examples. We also discuss some other examples in Appendix B, such as those involving multiple cooperating or competing agents, structured rejection sets, scientific meta-analysis, etc.

6 Constructing e-variables for sub-Gaussian rewards

We discuss the sub-Gaussian setting that has been the focus of existing methodology by JJ on bandit multiple testing. We explicitly define e-variables and an exploration component $(\mathcal{A}_t)$ that will have sample complexity bounds that match those of the algorithm derived by JJ, which uses p-variables. Specifically, we will consider the standard bandit setting where $|\mathcal{I}_t|=1$ and ν_i is 1-sub-Gaussian for each $i\in[k]$. Denote the means of each arm $i\in[k]$ as $\mu_i=\mathbb{E}[X_{i,t}]$ for all $t\in\mathbb{N}$. The goal is to find many arms where $\mu_i>\mu_0$, where we set $\mu_0=0$ to be the mean of a reward distribution under the null hypothesis. Thus, we define $\mathcal{H}_0=\{i\in[k]:\mu_i\leq\mu_0\}$ and $\mathcal{H}_1=\{i\in[k]:\mu_i>\mu_0\}$. Our framework ensures that $\mathrm{FDR}(\mathcal{S})\leq\delta$. Hence, we want algorithms that have $\mathrm{TPR}(\mathcal{S})\geq 1-\delta$ with small sample complexity. We include proofs of the results in this section in Appendix A.2.

We define the **predictably-mixed Hoeffding (PM-H)** e-process, after the corresponding confidence sequence defined in Waudby-Smith and Ramdas (2021), as follows:

$$E_{i,t}^{\text{PM-H}}(\mu_0)=\prod_{j=1}^{T_i(t)}\exp(\lambda_{i,t_i(j)}(X_{i,t_i(j)}-\mu_0)-\lambda_{i,t_i(j)}^2/2),$$

where $(\lambda_{i,t})$ is any sequence of nonnegative real numbers that is predictable w.r.t. $(\mathcal{F}_t)$.

Proposition 5. $E_{i,t}^{\text{PM-H}}(\mu_0)$ *is a nonnegative supermartingales, and thus an e-process, if $i\in\mathcal{H}_0$.*

Denote $\Delta_i \equiv \mu_i - \mu_0$ for $i \in \mathcal{H}_1$ and $\Delta \equiv \min_{i\in\mathcal{H}_1} \Delta_i$. Otherwise, let $\Delta_i \equiv \min_{j\in\mathcal{H}_1} \mu_i - \mu_0 = \Delta + (\mu_i - \mu_0)$. First, we recall a time-uniform bound on the sample mean $\widehat{\mu}_t$.

Fact 4 (Jamieson and Jain 2018). *Let $X_1, X_2, \ldots$ be i.i.d. draws from a 1-sub-Gaussian distribution with mean μ. The boundary $\varphi(t,\delta) = \sqrt{4\log(\log_2(2t)/\delta)/t}$ satisfies $\mathbb{P}(\exists t \in \mathbb{N} : |\widehat{\mu}_t - \mu| > \varphi(t,\delta)) \leq \delta$.*

$\varphi(t,\delta)$ is a time-uniform boundary, yielding a confidence sequence for the mean; tighter constants can be used, and such bounds can be derived under several nonparametric conditions (Howard et al., 2021). Thus, denote the sample mean at time t of each arm $i \in [k]$ by $\widehat{\mu}_{i,t}$. We formalize algorithm of JJ in terms of Algorithm 1 that utilizes φ from Fact 4 in (3). I_t is the single arm that the upper confidence bound (UCB) algorithm (3a) samples at time t. JJ prove the following sample complexity guarantee for the algorithm in (3).

$$I_t = \text{argmax}_{i\in[k]} \widehat{\mu}_{i,t-1} + \varphi(T_i(t),\delta), \tag{3a}$$

$$P_{i,t}^{\mathrm{JJ}} \equiv \inf_{\beta\in[0,1]} |\widehat{\mu}_{i,t} - \mu_0| > \varphi(t,\beta). \tag{3b}$$

Fact 5 (Jamieson and Jain 2018). *Let $(\mathcal{A}_t)$ output $\mathcal{I}_t = \{I_t\}$ and let $P_{i,t} = P_{i,t}^{\mathrm{JJ}}$ for all $i \in [k]$ and $t \in \mathbb{N}$ where I_t and $P_{i,t}^{\mathrm{JJ}}$ are defined in* (3). *Then, for any stopping time τ, Algorithm 1 will always guarantee* $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$. *With* $1-\delta$ *probability, there will exist* $T \lesssim \left(\sum_{i=1}^k \Delta_i^{-2} \log\log \Delta_i^{-2} + \Delta_i^{-2} \log(n/\delta)\right) \wedge k\Delta^{-2} \log(\log(\Delta^{-2})/\delta)$ *such that* $\mathrm{TPR}(\mathcal{S}_t) \geq 1-\delta$ *for all* $t \geq T$.

We formulate our e-variable algorithm in (4), where $R \in \mathbb{N}$ is a parameter of the algorithm. We discuss how we derived this choice of $(\lambda_{i,t})$ from betting strategies in Appendix D. We note that $E_{i,t}^{\text{PM-H}} / E_{i,R}^{\text{PM-H}}$ remains a e-process, since we are simply picking R as a new starting time for the e-variable to start counting samples. Essentially, the algorithm formulated in (4) has a warm start for R rounds where sample are accumulated on each arm so that $\widehat{\mu}_{i,t}$ becomes a better estimate of μ_i. Then, it launches into a UCB algorithm while using $E^{\text{PM-H}}$ as its e-variable. We prove the following guarantee.

$$I_t = \begin{cases} t \bmod k & t \leq R \\ \text{argmax}_{i\in[k]} \widehat{\mu}_{i,t-1} + \varphi(T_i(t-1),\delta) & t > R \end{cases}, \tag{4a}$$

$$\lambda_{i,t} = \widehat{\mu}_{i,t-1}/2 \vee 0, \qquad E_{i,t} = \begin{cases} 1 & t \leq R \\ \frac{E_{i,t}^{\text{PM-H}}}{E_{i,R}^{\text{PM-H}}} & t > R. \end{cases} \tag{4b}$$

Theorem 1. *Let $(\mathcal{A}_t)$ be such that $\mathcal{A}_t$ outputs $\mathcal{I}_t = \{I_t\}$ for all $t \in \mathbb{N}$, and I_t, $\lambda_{i,t}$, and $E_{i,t}$ are defined in* (4). *Let* $R = c_u k \Delta^{-2} \log(\log(\Delta^{-2})/\delta)$ *for some universal constant c_u that is agnostic of k, δ, and $\Delta_1, \ldots, \Delta_k$. Then, for any stopping time τ, Algorithm 1 will always guarantee* $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$. *With* $1-\delta$ *probability, there will exist with* $T \lesssim k\Delta^{-2} \log(\log(\Delta^{-2})/\delta)$ *such that* $\mathrm{TPR}(\mathcal{S}_t) \geq 1-\delta$ *for all* $t \geq T$.

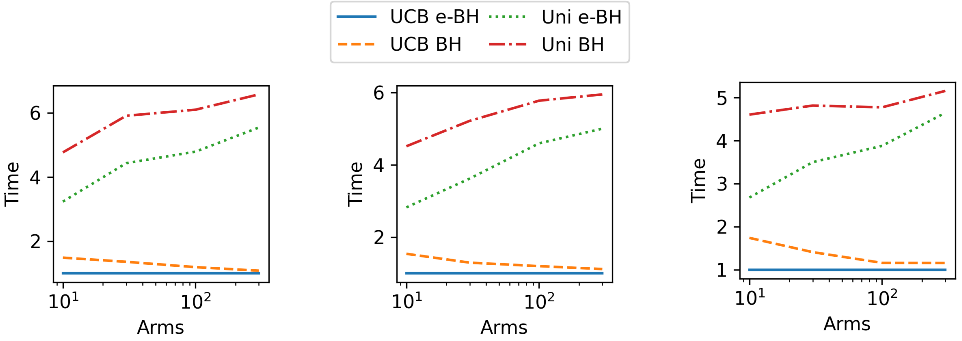
This matches the sample complexity of Fact 5 in the "similar means" case when the bound depends solely on Δ, that is all μ_i are similar, e.g. within constant factors of each other so that Δ is asymptotically equivalent to Δ_i for all $i \in \mathcal{H}_1$. This result can likely be generalized to exactly match Fact 5.

7 Numerical Simulations

We perform simulations for the situation discussed in Section 6 to demonstrate that our version of Algorithm 1 using e-variables is empirically as efficient as the algorithm of JJ, which uses p-variables. However, unlike JJ, our algorithm does not need to use a corrected level δ' that is based upon the dependence assumptions among $X_{1,t}, \ldots, X_{k,t}$ to guarantee FDR is controlled under δ. We explore additional simulations of combinatorial semi-bandit settings with dependent $X_{1,t}, \ldots, X_{k,t}$ in Appendix C that show the benefit of using e-variables over p-variables in our framework.

Simulation setup Let $\nu_i = \mathcal{N}(\mu_i, 1)$ where $\mu_i = \mu_0 = 0$ if $i \in \mathcal{H}_0$ and $\mu_i = 1/2$ if $i \in \mathcal{H}_1$. We consider 3 setups, where we set the number of non-null hypotheses to be $|\mathcal{H}_1| = 2, \log k$, and $\sqrt{k}$, so we may see the effect of different magnitudes of non-null hypotheses on the sample complexity of each method. We set $\delta = 0.05$ and compare 4 different methods. We compare the same two different exploration components for both e-variables and p-variables. The first exploration component we consider is simply uniform sampling across each arm (Uni). The second is the UCB sampling strategy described in (3a). When using BH, our choice for p-variables is the same as JJ and is formulated in (3b), and we set $\varphi(t,\delta) = \sqrt{\frac{2\log(1/\delta) + 6\log\log(1/\delta) + 3\log(\log(et/2))}{t}}$ in both the UCB method and our p-variable. When using e-BH, we set our evariables to $E_{i,t}^{\text{PM-H}}$ with $\lambda_{i,t} = \sqrt{\frac{2\log(2/\alpha)}{T_i(t)\log(T_i(t)+1)}}$, which is the default choice of $\lambda_{i,t}$ suggested in Waudby-Smith and Ramdas (2021). Recall that this choice of e-variables remains a e-process by Proposition 5 and thus maintains provable FDR control.

Results We observe that for uniform sampling, e-BH and e-variables seem to outperform BH and p-variables, although by a decreasing margin for more arms, especially in the case where $|\mathcal{H}_1| = \lfloor \sqrt{k} \rfloor$. For the UCB sampling algorithm, we see that e-variables and p-variables have relatively similar performance, with the gap narrowing as the number of arms increase as well. Thus, e-variables and e-BH empirically perform on par or better than p-variables with regards to sample complexity. This shows that using e-variables does not require any sacrifice in performance in simple cases where p-variables also work well. Further, e-variables do not require the same $\log k$ correction that p-variables need for situations where $X_{1,t},\ldots,X_{k,t}$ are arbitrarily dependent to guarantee FDR control at the same level. Thus, e-variables are preferable to p-variables as they are more flexible w.r.t. assumptions.



(a) $|\mathcal{H}_1| = 2$ (b) $|\mathcal{H}_1| = \lfloor \log k \rfloor$ (c) $|\mathcal{H}_1| = \lfloor \sqrt{k} \rfloor$

Figure 2: Relative comparison of time t for each method to obtain a rejection set, $\mathcal{S}_t$, that has a $\mathrm{TPR}(\mathcal{S}_t) \geq 1 - \delta$ while maintaining $\mathrm{FDR}(\mathcal{S}_t) \leq \delta$, where we choose $\delta = 0.05$. This plot compares e-BH vs. normal BH methods for evidence when using uniform sampling (Uni) or a UCB strategy for arm selection over different numbers of arms (choices of k) and different densities of non-null hypotheses (sizes of $\mathcal{H}_1$). Time is reported as a ratio to the time taken by UCB e-BH method. Note that e-variables perform on par or better than p-variables for both sampling strategies.

8 Conclusion, limitations and broader impact

In this paper, we developed a unified framework for bandit multiple hypothesis testing. We demonstrated that using the e-BH procedure applied to stopped e-processes can guarantee FDR control without any assumptions on the the dependency structure of $X_{1,t},\ldots,X_{k,t}$, the exploration strategy, or the stopping time of the algorithm, the presence of multiple agents, the ability of query multiple arms, etc. In contrast, existing algorithms using BH and p-variables have different FDR guarantees vary with the problem setting and dependence structure among the p-variables. In general, we argued that control of the FDR by BH can blow up by a factor of $\log k \log(1/\alpha)$, which means that the BH algorithm has to increase its threshold for discovery by $\log k \log(1/\alpha)$ to maintain FDR at the desired level. We provide more detailed explanations of these observations in the appendices. Despite its generality, we showed that in the standard sub-Gaussian reward setting, the instantiated algorithm matches sample complexity bounds with the p-variable algorithm by JJ for bounding the TPR, and has better practical performance despite improving JJ's guarantees by invoking the compliance results of Su (2018).

The appendices have additional examples of problem settings and simulations that show the utility of e-processes and our general framework. In fact, we can address an even more general setting where the null hypotheses do not have a one-to-one correspondence with the arms; in other words, despite the queries being at the arm-level, the hypotheses being tested could combine arms (for example, comparing different arms). We avoided this in the main paper for simplicity of exposition, since there were enough generalizations to describe in the simpler setup already.

The main limitation of the work is that it does not develop instance optimal sampling algorithms for multiple testing problem in the described settings with more complicated dependence structures; we believe this is a difficult open problem, requiring specialized techniques in each example. We do not foresee any negative societal impact of this work; it is aimed at reducing costs and improving reproducibility in scientific experimentation by controlling false discoveries in adaptive testing.

References

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010.

Y. Benjamini and Y. Hochberg. 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, 1995. ISSN 0035-9246.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, Aug. 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013699998.

L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly Optimal Sampling Algorithms for Combinatorial Pure Exploration. In *Conference on Learning Theory*, pages 482–534. PMLR, June 2017.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. *Neural Information Processing Systems*, 2014.

W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

P. Grünwald, R. de Heide, and W. Koolen. Safe Testing. *arXiv:1906.07801 [cs, math, stat]*, June 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, et al. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

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

L. Jain and K. G. Jamieson. A new perspective on pool-based active classification and false-discovery control. In *Neural Information Processing Systems*, 2019.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. Lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, May 2014.

K. G. Jamieson and L. Jain. A bandit approach to sequential experimental design with false discovery control. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

M. Jourdan, M. Mutný, J. Kirschner, and A. Krause. Efficient Pure Exploration for Combinatorial Bandits with Semi-Bandit Feedback. In *Algorithmic Learning Theory*, pages 805–849. PMLR, Mar. 2021.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 227–234, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, Jan. 2016. ISSN 1532-4435.

J. L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4): 917–926, 1956. doi: 10.1002/j.1538-7305.1956.tb03809.x.

L. Lei, A. Ramdas, and W. Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, July 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa064.

X. Nie, X. Tian, J. Taylor, and J. Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2018.

A. Ramdas, J. Chen, M. J. Wainwright, and M. I. Jordan. A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika*, 106:69–86, Mar. 2019. ISSN 0006-3444. doi: 10.1093/biomet/asy066.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167 [math, stat]*, Sept. 2020.

G. Shafer. The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A (forthcoming discussion paper)*, 2020.

J. Shin, A. Ramdas, and A. Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *Neural Information Processing Systems*, 2019.

J. Shin, A. Ramdas, and A. Rinaldo. On conditional versus marginal bias in multi-armed bandits. In *International Conference on Machine Learning*, pages 8852–8861. PMLR, 2020.

J. Shin, A. Ramdas, and A. Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*, 2021.

W. J. Su. The FDR-Linking Theorem. *arXiv:1812.08965 [math, stat]*, Dec. 2018.

M. Valko. *Bandits on Graphs and Structures*. HdR Thesis, Ecole normale supérieure de Paris-Saclay, Paris-Saclay, France, June 2016.

V. Vovk and R. Wang. True and false discoveries with e-values. *arXiv:1912.13292 [math, stat]*, Feb. 2020.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *Annals of Statistics (forthcoming)*, 2021.

R. Wang and A. Ramdas. False discovery rate control with e-values. *arXiv:2009.02824 [math, stat]*, Nov. 2020.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv:2010.09686 [math, stat]*, Feb. 2021.

F. Yang, A. Ramdas, K. Jamieson, and M. J. Wainwright. A framework for multi-A(rmed)/B(andit) testing with online FDR control. *Neural Information Processing Systems*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4 for framework, Section 5 for differing dependence guarantees, Section 6 for theoretical matching results and Section 7 for matching empirical results.
   (b) Did you describe the limitations of your work? [Yes] See Section 8
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Problem statement is specified in Section 1 under "Formal bandit setting". Additional setting assumptions are specified in Section 5 and Section 6.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplement
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Configs specifying all parameters of each experiment
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Error bars are visually small compared to plot size.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Only personal laptop utilized.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]