A unified framework for bandit multiple testing

Ziyu Xu[1], Ruodu Wang[3], Aaditya Ramdas[1,2]

Departments of [1]Statistics and [2]Machine Learning, Carnegie Mellon University
[3]Department of Statistics and Actuarial Science, University of Waterloo

{xzy,aramdas}@cmu.edu, wang@uwaterloo.ca

November 18, 2021

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.

Contents

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 typically 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 drug 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. [46], one 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 [19], henceforth called JJ, but our techniques apply equally well to the above setup. *To avoid confusions, note that our setup is very different from the active classification work of the same authors [17].* 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 (potentially composite) null hypothesis — for example, the hypothesis that 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 (at most) zero for null arms and positive for non-nulls. Thus, a strategy 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 D.3.

Our contribution. We introduce a modular meta-algorithm for bandit multiple testing with provable FDR control that utilizes "e-values" — or, more appropriately, their sequential analog, "e-processes" — a recently introduced alternative to p-values (or p-processes) by Ramdas et al. [32] for various testing problems, that are inherently related to martingales, gambling and betting [33, 14, 15, 44]. This work is the first to carefully study e-processes in general MAB settings, building on prior work that studied a special case [44]. We also are the first to extend the bandit multiple testing problem to the combinatorial bandit setting — JJ had previously only analyzed the problem in the single-arm, independent reward setting. 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 D.3). (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. Consequently, FDR control in our meta-algorithm is agnostic to much of problem setup and can be proved in a vast array of settings. This is not true when working for p-values. In particular, the techniques for proving FDR control in JJ are highly reliant on the specific bandit setup in their paper. (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 C based on best arm identification algorithms [1, 22, 11, 18, 23, 10, 20].

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 [6] 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 [44], 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 problem setup. We define the bandit as having k arms, and ν_i as the (unknown) reward distribution for arm $i \in [k] = \{1, \dots, k\}$. 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". While we assume by default that all rewards from an arm are i.i.d., we also formulate tests for hypotheses on reward distributions that may violate this assumption in Appendix G. 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 $\mathcal{H}_1 := [k] \setminus \mathcal{H}_0$.

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 $\mathcal{I}_t \in \mathcal{K}$ to sample jointly from. The special choice of $\mathcal{K} = \{\{1\}, \{2\}, \dots, \{k\}\}$ recovers the standard bandit setup, but otherwise this setting is known as combinatorial bandits with semi-bandit feedback [12]. We also consider the special case of full-bandit feedback (the algorithm sees all rewards at each time step) in Appendix D.1. 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 $t \in \mathbb{N}$: $\mathcal{F}_t := \sigma\left(U \cup \{(i, s, X_{i,j}) : s \leq t, i \in \mathcal{I}_s\}\right)$ and we let $\mathcal{F}_0 := \sigma(U)$ 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 to 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. These two metrics are defined as follows:

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

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. On the other hand, TPR is a measurement of the power of the algorithm i.e. how many of the non-null hypotheses does the algorithm discover. Our implicit goal in the multiple testing problem is to maximize the number of true discoveries while not making too many mistakes i.e. keep the FDR controlled.

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. For a

particular setting where we make such a set of assumptions, we demonstrate in Section 4 that we can prove TPR guarantees for algorithms within our framework. Hence, our FDR controlling framework is not vacuous as it includes powerful algorithms i.e. algorithms which make many true discoveries. However, our focus is primarily to show that the FDR control of our framework is robust to a wide range of conditions.

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 $\text{FDR}(\mathcal{S}_{\tau^*}) \leq \delta$ and large TPR as quickly as possible.

Now that we have specified the problem we are interested in, we can introduce our main technical tools for ensuring FDR control at stopping times: e-processes and p-processes.

2 Technical preliminaries

2.1 E-processes versus p-processes

An 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 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 to 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 $\sup_{\tau\in\mathcal{T}} \mathbb{E}[E_\tau] \leq 1$ (let $E_\infty := \limsup_{t\in\mathbb{N}} E_t$ and $P_\infty := \liminf_{t\in\mathbb{N}} P_t$). These sequentially valid forms of p-variables and e-variables are crucial since we allow the bandit algorithm to stop and output a rejection set in a data-dependent manner. 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 [31], but most importantly, nonnegative (super)martingales play a central role 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 [45] are also important for the development of sample efficient algorithms and we discuss how we use betting ideas in Appendix F. We also discuss some useful equivalence properties of p-processes in Appendix A.1, while Appendix A.2 introduces supermartingales for the unfamiliar reader.

Why use e-processes over p-processes? Wang and Ramdas [44] 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 both better flexibility and sample complexity of the algorithm. While this question has been the focus of a recent line of work for hypothesis tests in general [33, 41, 14, 44], 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 be robust to arbitrary dependencies between statistics computed for each arm without additional correction. Further, we explore how e-processes can be merged under different conditions in Appendix D.3 to facilitate incorporation of existing evidence and cooperation between multiple agents and present concrete ways to construct e-processes in Appendices B.3 and F.

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 [27, 35, 36, 37]. 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, e-variable based algorithms enable us to prove FDR guarantees without assumptions on the sampling method.

In contrast, procedures involving p-variables, such as the ones used in JJ, require the test level of α to be corrected by a factor of at least $\log(1/\alpha)$ when rewards are independent across arms, and a factor of $\log k$ otherwise. We expand on this in Section 2.2.

2.2 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 [6], 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 the degree of FDR guarantee (i.e. test level).

Benjamini-Hochberg (BH) requires corrections for dependence and self-consistency. A set $\mathcal{S}$ of p-values is called *p-self-consistent* [9] at level α iff:

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

The BH procedure with input $p_1, \ldots, p_k$ outputs the largest p-self-consistent set w.r.t. the input, which we denote $\text{BH}[\alpha](p_1, \ldots, p_k)$. 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). A formal definition is provided in Benjamini and Yekutieli [7], and it is sufficient for our purposes 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 [6], Benjamini and Yekutieli [7])**.** *Let* $\mathcal{S} = \text{BH}[\alpha](p_1, \ldots, p_k)$. *If* $P_1, \ldots P_k$ *are PRDS, then* $\text{FDR}(\mathcal{S}) \leq \alpha$. *Otherwise, under arbitrary dependence amongst* $P_1, \ldots P_k$, *the BH procedure ensures* $\text{FDR}(\mathcal{S}) \leq \alpha\ell_k$, *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 $\ell_k \approx \log k$. A larger FDR guarantee is provided for arbitrary p-self-consistent sets.

Fact 2 (P-self-consistent FDR control. Su [38], Blanchard and Roquain [9], Wang and Ramdas [44])**.** *If* $\mathcal{S}$ *is p-self-consistent at level* α *and* $P_1, \ldots, P_k$ *satisfy PRDS,* [1] *then* $\text{FDR}(\mathcal{S}) \leq \alpha(1 + \log(1/\alpha))$. *Otherwise, when there is arbitrary dependence among* $P_1, \ldots, P_k$, $\text{FDR}(\mathcal{S}) \leq \alpha\ell_k$ *(consequence of Propositions 2.7 and 3.7 from Blanchard and Roquain [9] and Proposition 5.2 from Wang and Ramdas [44]).*

These two facts do not imply each other; the BH procedure outputs the largest self-consistent set and has a stronger or equivalent error guarantee under either type of dependence. While it may seem like we should always use BH and the guarantee from Fact 1 to form a rejection set, we elaborate in Section 3.1 on how we can use Fact 2 to provide FDR control for BH when the p-variables are not necessarily PRDS, and in settings where we may not directly use BH.

e-BH needs no correction for dependence or self-consistency. The e-BH procedure created by Wang and Ramdas [44] uses e-variables instead of p-variables and proceeds similarly to the BH procedure. In this case, let $e_1, \ldots, e_k$ be the realized e-values for a set of e-variables $E_1, \ldots, E_k$. Define $e_{[i]}$ to be the ith largest e-value for $i \in [k]$. A set $\mathcal{S}$ is *e-self-consistent* at level α iff $\mathcal{S}$ satisfies the following:

$$\min_{i \in \mathcal{S}} \; e_i \geq \frac{k}{\alpha|\mathcal{S}|}. \tag{2}$$

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

[1] Su [38] technically employs a *slightly* weaker condition which implies PRDS, and refers to self-consistency as "compliance" (or, better said, compliance is a special case of self-consistency).

Fact 3 (E-variable self-consistency FDR control. Wang and Ramdas [44]). *If $\mathcal{S}$ is e-self-consistent at level α, then $\mathrm{FDR}(\mathcal{S}) \leq \alpha$ regardless of the dependence structure.*

All FDR bounds discussed in Facts 1 to 3 are optimal, in the sense that there exist e-variable/p-variable distributions with an FDR that is arbitrarily close or equivalent to the stated bound. Consequently, e-variables are more advantageous, since their FDR control does not change under different types of dependence as opposed to the factor of $1 + \log(1/\alpha)$ or $\log k$ p-variables pay on the FDR for different settings.

In the case where p-variables can only be constructed as $P = 1/E$, where E is an e-variable, the rejection sets output by BH and e-BH are identical. However, the e-self-consistency guarantee in Fact 3 provides identical or tighter FDR control than the BH procedure guarantee in Fact 1 or p-self-consistency 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 $\mathrm{FDR}(\mathcal{S}) \leq \delta$ for a fixed δ under different dependence structures. We now provide a meta-algorithm that utilizes p-self-consistency and e-self-consistency to guarantee FDR control in the bandit setting.

3 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 τ^*, Let $(p_{1,t}), \ldots, (p_{k,t})$ and $(e_{1,t}), \ldots, (e_{k,t})$ denote the realized values of p-processes and e-processes, respectively. Let the desired level of FDR control be $\delta \in (0,1)$. Let δ' be the correction of δ for BH based upon the dependencies of $X_{1,t}, \ldots, X_{k,t}$. Set $D_0 = \emptyset$.

for t *in* $1 \ldots$ **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).
- $\mathcal{S}_t := \begin{cases} \mathrm{BH}[\delta'](p_{1,t}, \ldots, p_{k,t}) \text{ or arbitrary p-self-consistent set} & \text{if using p-variables} \\ \mathrm{eBH}[\delta](e_{1,t}, \ldots, e_{k,t}) \text{ or arbitrary e-self-consistent set} & \text{if using e-variables} \end{cases}$
- **if** $\tau^* = t$ **then** stop and **return** $\mathcal{S}_t$;

end

Exploration component. This 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 $D_t := \{(i, j, X_{i,j}) : j \leq t, i \in \mathcal{I}_j\}$ 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 $\mathrm{FDR}(\mathcal{S}) \leq \delta$ 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 formulation of the candidate rejection set, $\mathcal{S}_t \subseteq [k]$, at each time $t \in \mathbb{N}$ in the evidence component. This construction is completely separate from $(\mathcal{A}_t)$. Critically, $(\mathcal{S}_t)$ satisfies $\mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ for any stopping time $\tau \in \mathcal{T}$. This is accomplished by applying BH or e-BH to p-processes or e-processes, respectively. At stopping time τ, $P_{i,\tau}$ is a p-variable when $(P_{i,t})$ is a p-process, and similarly $E_{i,\tau}$ is an e-variable when $(E_{i,t})$ is an e-process. Thus, $\mathcal{S}_\tau$ is the result of applying BH to p-variables or e-BH to e-variables.

Consequently, the aforementioned framework allows us to guarantee $\sup_{\tau \in \mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ in a way that is agnostic to the exploration component. For completeness, we do discuss some generic exploration strategies in Appendix C. 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.

3.1 FDR control under different dependence structures

In the general combinatorial bandit setting, different dependence structures affect the choice of δ' that ensures FDR control at δ in the p-variable and BH case. Table 1 summarizes the guarantees and choices of δ' 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 $X_{1,t}, \dots, X_{k,t}$ are independent. Critically, JJ employ BH and p-variables in their algorithm, and the FDR guarantee of BH changes based on the dependencies between reward distributions. 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 $X_{1,t}, \dots, X_{k,t}$ 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 B.1.

Table 1: FDR control for 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 *any e-self-consistent procedure provides* $\text{FDR}(\mathcal{S}) \leq \alpha$ *in all settings in the table.*

	Dependence of $X_{1,t}, \dots, X_{k,t}$	
Adaptivity of $(\mathcal{A}_t)$ and τ^*	*independent*	*arbitrarily dependent*
non-adaptive	$\text{FDR}(\mathcal{S}) \leq \alpha$ $\delta' = \delta$	$\text{FDR}(\mathcal{S}) \leq \alpha \log k$ $\delta' = \delta / \log k$ (Prop. 2)
adaptive	$\text{FDR}(\mathcal{S}) \leq \alpha((1 + \log(1/\alpha)) \wedge \log k)$ $\delta' = c_\delta \vee (\delta / \log k)$ (Prop. 1)	
Any **e-self-consistent procedure** ensures $\text{FDR}(\mathcal{S}) \leq \alpha$ in all settings and sets $\alpha = \delta$.		

Adaptive $(\mathcal{A}_t)$ and independent $X_{1,t}, \dots, X_{k,t}$. JJ consider this case in the non-combinatorial bandit setting, but their insights and techniques also can be extended to the combinatorial setting. We give a sketch of their proof here, and produce the full proof in Appendix B.1. In the language of self-consistency (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-self-consistent with a different set of independent p-variables.* Define $P_1^*, \dots, P_k^*$, 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 a p-process for each arm $i \in [k]$, the corresponding P_i^* is a p-variable (Proposition 6 in Appendix A.1). Further, $P_1^*, \dots, P_k^*$ are independent because $X_{1,t}, \dots, X_{k,t}$ are independent. By definition of $P_1^*, \dots, P_k^*$, $p_i^* \leq p_{i,t}$ for any $i \in [k]$ and any $t \in \mathbb{N}$. Thus, $\mathcal{S}_{\tau^*}$ is p-self-consistent w.r.t. $p_1^*, \dots, p_k^*$, and has its FDR bounded by $\alpha(1 + \log(1/\alpha))$ due to Fact 2. At the same time, the arbitrary dependence guarantee from Fact 1 still applies. Combining these facts, we achieve the following guarantee:

Proposition 1. *When $(\mathcal{A}_t)$ is adaptive and $X_{1,t}, \dots, X_{k,t}$ are independent, Algorithm 1 with p-processes and an arbitrary p-self-consistent set guarantees $\sup_{\tau \in \mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \leq \delta$ if $\delta' \leq c_\delta \vee \delta/\ell_k$, where for any $\delta \in (0, 1)$, define $c_\delta \leq \delta$ as the solution to $c_\delta(1 + \log(1/c_\delta)) = \delta$.*

Note that Proposition 1 is valid for any p-self-consistent set since p-self-consistency is the only property required of the output set to prove the result. JJ prove a similar bound to Proposition 1. However, they used a larger FDR bound for p-self-consistent sets with worse constants (which was subsequently improved by Su [38] as presented earlier), and they only considered the non-combinatorial case. Proposition 1 uses an optimal bound on p-self-consistent sets from Fact 2, and is valid in our combinatorial bandit setup.

Adaptive $(\mathcal{A}_t)$ and arbitrarily dependent $X_{1,t}, \dots, X_{k,t}$. 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 just 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.

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

Finally, consider a setting structured setting where we cannot output the rejection set of BH. Such a constraint often occurs in directed acyclic graph (DAG) settings where there is a hierarchy among hypotheses that restricts which rejection sets are allowed [30, 25]. Instead, we would like to output the largest self-consistent set that respects the structural constraints. By Fact 2, we get the following FDR control.

Proposition 3. *If $(\mathcal{A}_t)$ is adaptive and $X_{1,t}, \ldots, X_{k,t}$ are dependent, Algorithm 1 with p-variables that outputs an arbitrary p-self-consistent $\mathcal{S}_t$ guarantees* $\sup_{\tau \in \mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ *if* $\delta' \leq c_\delta/\ell_k$.

We explore the structured setting with greater depth in Appendix D.2. Unlike p-variables, e-variables do not need correction in any of the aforementioned settings.

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

Thus, running e-BH (or any e-self-consistent procedure) at level δ is valid for any choice of $(\mathcal{A}_t)$ and type of dependence. Now, we give an example where $X_{1,t}, \ldots, X_{k,t}$ might be arbitrarily dependent.

3.2 Illustrative examples to demonstrate flexibility of the framework

Below, we briefly describe a set of nontrivial illustrative examples to showcase the flexibility of our framework. In most of the cases below, a p-process approach would have to correct for dependence and/or self-consistency in different case-specific ways, rendering it more conservative and requiring careful arguments to justify FDR control. However, working with our unified framework is easy, handling both self-consistency and dependence issues in the same breath and without any changes to the algorithm or analysis. The data scientist can focus on designing powerful e-processes *for each arm separately* and let the modular framework correct for the multiplicity aspect.

Example: sampling nodes on a graph. A scenario where $X_{1,t}, \ldots, X_{k,t}$ 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 [26] and have many real world applications [39]. 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, and this data can be arbitrarily dependent across the sensors. Figure 1 illustrates a superarm in this situation.

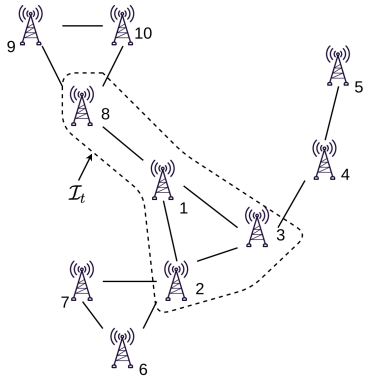


Figure 1: A superarm consists of a node and all its neighbors. The dotted line captures $\mathcal{I}_t$, the superarm around node 1.

In such a setting, if Proposition 2 is used to guarantee $\text{FDR}(\mathcal{S}) \leq \delta$ 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 E.2, and show these differences empirically. We also discuss some other examples in the appendix that we will summarize here.

- **Multiple agents** (Appendix D.3): Consider the setting where multiple agents are operating on the same bandit, and we want to aggregate the evidence for rejection across agents. For e-processes, we present an algorithm for merging e-values that maintains FDR control.
- **Structured rejection sets** (Appendix D.2): We illustrate the difference between self-consistency guarantees for p-variables and e-variables when a DAG hierarchy is imposed upon the hypotheses.
- **Multi-arm hypotheses** (Appendix D.4) A hypothesis may concern the reward distributions of multiple arms e.g. are the means of two different arms equivalent? We provide FDR guarantees even when hypotheses and arms are not matched one-to-one.
- **Streaming data setting** (Appendix D.1) Our methods also naturally extend to the streaming setting when the algorithm views the rewards of every at each time step.

Now that we have shown FDR is controlled using e-variables in a way that is robust to the underlying dependence structure, we analyze the sample complexity of achieving a high TPR using e-variables when the rewards are independent and sub-Gaussian.

4 E-process sample complexity guarantees for sub-Gaussian arms

We provide sample complexity guarantees for the sub-Gaussian setting that has been the focus of existing methodology by JJ in bandit multiple testing. We explicitly define e-processes and an exploration component $(\mathcal{A}_t)$ that will have sample complexity bounds matching those of the algorithm in 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 $\text{FDR}(\mathcal{S}) \leq \delta$, and we also want to achieve $\text{TPR}(\mathcal{S}) \geq 1 - \delta$ with small sample complexity. Proofs of the results from this section are in Appendices B.2 and B.3. As an aside, we also discuss what hypotheses we can test when the reward distribution is not necessarily independent across $t \in \mathbb{N}$, but the conditional distribution of the rewards still satisfy certain sub-Gaussian guarantees in Appendix G.

Our e-process of choice is the **discrete mixture e-process** from Howard et al. [16]:

$$E_{i,t}^{\text{DM}}(\mu_0) := \sum_{\ell=0}^{\infty} w_\ell \exp\left(\sum_{j=1}^{T_i(t)} \lambda_\ell (X_{i,t_i(j)} - \mu_0) - \lambda_\ell^2/2\right), \tag{3a}$$

$$\text{where } \lambda_\ell := \tfrac{1}{e^{\ell+5/2}} \text{ and } w_\ell := \tfrac{2(e-1)}{e(\ell+2)^2} \text{ for } \ell \in \mathbb{N}_0. \tag{3b}$$

Proposition 5. *$E_{i,t}^{\text{DM}}$ is an e-process when ν_i is 1-sub-Gaussian and $i \in \mathcal{H}_0$.*

Denote $\Delta_i \equiv \mu_i - \mu_0$ for $i \in \mathcal{H}_1$ and $\Delta \equiv \min_{i \in \mathcal{H}_1} \Delta_i$. When $i \in \mathcal{H}_0$, 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 (JJ, Kaufmann et al. [23], Howard et al. [16])**.** *Let $X_1, X_2, \ldots$ be i.i.d. draws from a 1-sub-Gaussian distribution with mean μ. Consider the boundaries defined in (4). Let φ be one of these boundaries. Then,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : |\widehat{\mu}_t - \mu| > \varphi(t, \delta)\right) \leq \delta$$

for any $\delta \in (0, 1)$ if $\varphi \in \{\varphi^0, \varphi^{\text{IS}}\}$ and any $\delta \in (0, 0.1]$ if $\varphi = \varphi^{\text{JJ}}$.

$$\varphi^0(t,\delta) \coloneqq \sqrt{\tfrac{4\log(\log_2(2t)/\delta)}{t}}, \tag{4a}$$

$$\varphi^{\mathrm{JJ}}(t,\delta) \coloneqq \sqrt{\tfrac{2\log(1/\delta)+6\log\log(1/\delta)+3\log(\log(et/2))}{t}}, \tag{4b}$$

$$\varphi^{\mathrm{IS}}(t,\delta) \coloneqq \sqrt{\tfrac{2.89\log\log(2.041t)+2.065\log\left(\frac{4.983}{\delta}\right)}{t}}. \tag{4c}$$

We will use φ to refer to an arbitrary boundary from Fact 4. All of the φ are time-uniform boundaries that yield confidence sequences for the mean. Note that φ^0 is generally larger than the other boundaries, so we use φ^0 as the default boundary in our proofs, and we explore how different choices of φ affect empirical performance in Appendix E.1. Now, we can define the algorithm from JJ in (5), which consists of an exploration policy based on an upper confidence bound (UCB) of the mean reward (specified by a singleton set $\mathcal{I}_t = \{I_t\}$) and a p-variable derived from Fact 4.

In (5a), we denote the sample mean at time t of each arm $i \in [k]$ by $\widehat{\mu}_{i,t}$. Let $f \lesssim g$ denote f asymptotically dominates g i.e. there exist $c > 0$ that is independent of the problem parameters such that $f \leq cg$. JJ prove the following sample complexity guarantee for their algorithm.

$$I_t = \mathrm{argmax}_{i\in[k]\setminus\mathcal{S}_{t-1}} \widehat{\mu}_{i,t-1} + \varphi(T_i(t-1),\delta), \tag{5a}$$

$$P_{i,t} \equiv \inf\{\rho \in [0,1] : |\widehat{\mu}_{i,t} - \mu_0| > \varphi(t,\rho)\}. \tag{5b}$$

Fact 5 (From JJ). *Let $(\mathcal{A}_t)$ output $\mathcal{I}_t = \{I_t\}$, and let I_t and $P_{i,t}$ be specified by Alg. 5 with $\varphi = \varphi^0$. Then, Algorithm 1 will always guarantee $\sup_{\tau\in\mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$. With at least $1-\delta$ probability, there will exist*

$$T \lesssim \left(\sum_{i=1}^{k} \Delta_i^{-2}\log\log\Delta_i^{-2} + \Delta_i^{-2}\log(k/\delta)\right) \wedge k\Delta^{-2}\log(\log(\Delta^{-2})/\delta)$$

such that $\mathrm{TPR}(\mathcal{S}_t) \geq 1-\delta$ for all $t \geq T$.

We show that we can match the sample complexity bounds of Fact 5 with e-variables.

Theorem 1. *Let ν_i be 1-sub-Gaussian for $i \in [k]$. Set $(\mathcal{A}_t)$ so $\mathcal{A}_t$ outputs $\{I_t\}$ from (5a) for all $t \in \mathbb{N}$ and $E_{i,t}$ to $E_{i,t}^{\mathrm{DM}}$. Algorithm 1 ensures $\sup_{\tau\in\mathcal{T}} \mathrm{FDR}(\mathcal{S}_\tau) \leq \delta$ and, with at least $1-\delta$ probability, there exists*

$$T \lesssim \left(\sum_{i=1}^{k} \Delta_i^{-2}\log\log\Delta_i^{-2} + \Delta_i^{-2}\log(k/\delta)\right) \wedge k\Delta^{-2}\log(\log(\Delta^{-2})/\delta)$$

such that $\mathrm{TPR}(\mathcal{S}_t) \geq 1-\delta$ for all $t \geq T$.

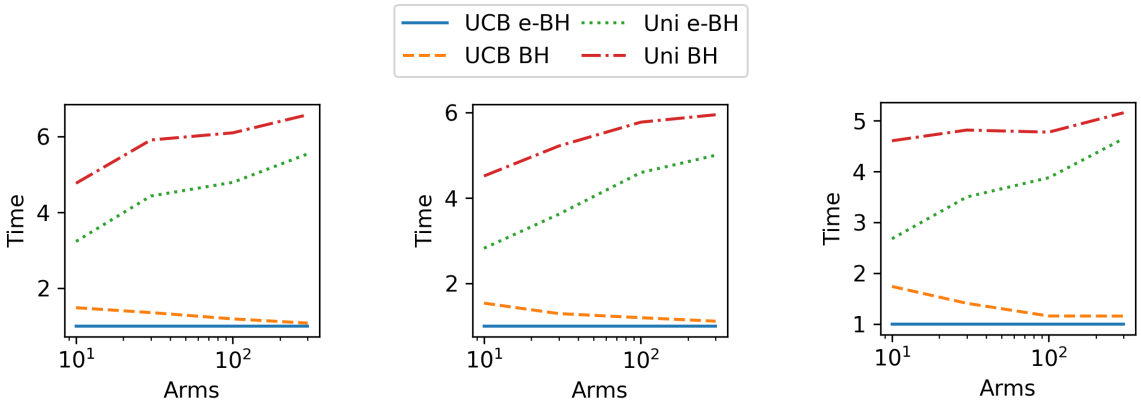
In addition to matching theoretical guarantees, we show in the following section that e-variables and e-BH perform empirically as well or better than p-variables and BH through numerical simulations.

5 Numerical simulations

We perform simulations for the sub-Gaussian setting discussed in Section 4 to demonstrate that our version of Algorithm 1 using e-variables is empirically as efficient as the algorithm of JJ, which uses p-variables (code available here) . However, unlike JJ, our algorithm does not use a corrected level δ' based upon the dependence assumptions among $X_{1,t},\ldots,X_{k,t}$ to guarantee FDR is controlled at level δ. We explore additional simulations of combinatorial semi-bandit settings with dependent $X_{1,t},\ldots,X_{k,t}$ in Appendix E 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 $\sqrt{k}$, to 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 (5a). When using BH, our formulation for p-variables is (5b), which is the same as JJ. Like JJ, we set $\varphi = \varphi^{\text{JJ}}$ in our simulations. When using e-BH, we set our e-variables to $E_{i,t}^{\text{PM-H}} := \prod_{j=1}^{T_i(t)} \exp(\lambda_{i,t_i(j)}(X_{i,t_i(j)} - \mu_0) - \lambda_{i,t_i(j)}^2/2)$ 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 [45]. We show that this is a valid e-process in Appendix F and maintains FDR control.

Results We plot the relative performance of each method to e-BH with UCB sampling in Figure 2. 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 to obtain a rejection set, $\mathcal{S}_t$, that has a $\text{TPR}(\mathcal{S}_t) \geq 1 - \delta$ and $\text{FDR}(\mathcal{S}_t) \leq \delta$ where $\delta = 0.05$. This plot compares e-BH vs. BH for both uniform (Uni) and UCB sampling over different numbers of arms (choices of k) and 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 the methods using e-variables perform on par or better than methods using p-variables for both sampling strategies.

6 Conclusion

In this paper, we developed a unified framework for bandit multiple hypothesis testing. We demonstrated that applying the e-BH procedure to stopped e-processes guarantees FDR control without assumptions on the the dependency between $X_{1,t}, \ldots, X_{k,t}$, exploration strategy, stopping time of the algorithm, ability to query multiple arms, etc. In contrast, existing algorithms using BH and p-variables have FDR guarantees that vary with the problem setting and dependence structure among the p-variables. We argued that control of the FDR with p-variables can blow up by a factor of $\log k$, and any p-self-consistent algorithm must decrease its threshold for discovery correspondingly to maintain FDR control at the desired level. We provide more detailed explanations of these observations in Appendix D.2. In addition to demonstrating the generality of our meta-algorithm, we showed that in the standard sub-Gaussian reward setting, the instantiated algorithm

matches the sample complexity bounds of the p-variable algorithm by JJ for achieving high TPR, and has better practical performance than JJ's algorithm, despite the fact that we improve JJ's guarantees by invoking the self-consistency results of Su [38].

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 also discuss the multi-agent setting where there could be multiple agents that operate the same bandit. We avoided these scenarios in the main paper for simplicity of exposition, since there were enough generalizations to describe in the simpler setup already.

Acknowledgments RW acknowledges funding from NSERC RGPIN-2018-03823 and RGPAS-2018-522590. AR acknowledges funding from NSF DMS 1916320 and ARL IoBT REIGN. Research reported in this paper was sponsored in part by the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBTCRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010. 3, 23

[2] J. Bartroff. Multiple Hypothesis Tests Controlling Generalized Error Rates for Sequential Data. *Statistica Sinica*, 2017. ISSN 10170405. 24

[3] J. Bartroff and J. Song. Sequential tests of multiple hypotheses controlling type I and II familywise error rates. *Journal of Statistical Planning and Inference*, 153:100–114, 2014. ISSN 0378-3758. 24

[4] J. Bartroff and J. Song. A Rejection Principle for Sequential Tests of Multiple Hypotheses Controlling Familywise Error Rates. *Scandinavian Journal of Statistics*, 31(1):3–19, 2016. ISSN 0303-6898. 24

[5] J. Bartroff and J. Song. Sequential Tests of Multiple Hypotheses Controlling False Discovery and Nondiscovery Rates. *Sequential analysis*, 39(1):65–91, 2020. ISSN 0747-4946. 24

[6] 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. 3, 6

[7] 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, 2001. ISSN 0090-5364, 2168-8966. 6

[8] A. Birnbaum. Combining Independent Tests of Significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954. ISSN 0162-1459. 26

[9] G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008. ISSN 1935-7524, 1935-7524. 6

[10] L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly Optimal Sampling Algorithms for Combinatorial Pure Exploration. In *Conference on Learning Theory*, 2017. 3, 23

[11] S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. *Neural Information Processing Systems*, 2014. 3

[12] 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. 4, 23

[13] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5a edition, 2017. 16

[14] P. Grünwald, R. de Heide, and W. Koolen. Safe Testing. *arXiv:1906.07801*, 2020. 3, 5

[15] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. 3, 17

[16] 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. 10, 16, 17, 18

[17] L. Jain and K. G. Jamieson. A new perspective on pool-based active classification and false-discovery control. In *Neural Information Processing Systems*, 2019. 3

[18] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. Lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Conference on Learning Theory*, 2014. 3, 19, 23

[19] K. G. Jamieson and L. Jain. A bandit approach to sequential experimental design with false discovery control. In *Neural Information Processing Systems*, volume 31, 2018. 3

[20] M. Jourdan, M. Mutný, J. Kirschner, and A. Krause. Efficient Pure Exploration for Combinatorial Bandits with Semi-Bandit Feedback. In *Algorithmic Learning Theory*, 2021. 3, 23

[21] K.-S. Jun, F. Orabona, S. Wright, and R. Willett. Online learning for changing environments using coin betting. *Electronic Journal of Statistics*, 11(2):5282–5310, 2017. ISSN 1935-7524, 1935-7524. 31

[22] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2012. 3, 23

[23] 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, 2016. ISSN 1532-4435. 3, 10, 23

[24] J. L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956. 32

[25] L. Lei, A. Ramdas, and W. Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 2020. ISSN 0006-3444. 9, 24

[26] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Neural Information Processing Systems*, 2011. 9

[27] 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*, 2018. 5

[28] F. Orabona and D. Pál. Coin betting and parameter-free online learning. In *Neural Information Processing Systems*, 2016. 31

[29] F. Orabona and T. Tommasi. Training Deep Networks without Learning Rates Through Coin Betting. In *Neural Information Processing Systems*, 2017. 31

[30] 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, 2019. ISSN 0006-3444. 9, 24

[31] A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020. 5, 16, 28

[32] A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2021. ISSN 0888-613X. 3

[33] G. Shafer. 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, 2021. ISSN 1467-985X. 3, 5, 27, 30

[34] G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, Ltd, first edition, 2019. ISBN 978-1-118-54803-5. 30

[35] J. Shin, A. Ramdas, and A. Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *Neural Information Processing Systems*, 2019. 5

[36] J. Shin, A. Ramdas, and A. Rinaldo. On conditional versus marginal bias in multi-armed bandits. In *International Conference on Machine Learning*, 2020. 5

[37] J. Shin, A. Ramdas, and A. Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *SIAM Journal on the Mathematics of Data Science*, 2021. 5

[38] W. J. Su. The FDR-Linking Theorem. *arXiv:1812.08965*, 2018. 6, 8, 13

[39] M. Valko. *Bandits on Graphs and Structures*. HdR Thesis, Ecole normale supérieure de Paris-Saclay, Paris-Saclay, France, 2016. 9

[40] V. Vovk. Testing randomness online. *Statistical Science*, 2021. 28

[41] V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 2021. 5, 25

[42] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, and A. Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Symposium on Conformal and Probabilistic Prediction and Applications*, 2021. 28

[43] V. Vovk, B. Wang, and R. Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics (forthcoming)*, 2021. 26

[44] R. Wang and A. Ramdas. False discovery rate control with e-values. *arXiv:2009.02824*, 2020. 3, 4, 5, 6, 7

[45] I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv:2010.09686*, 2021. 5, 12, 31, 38

[46] 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. 2

A Miscellaneous technicalities

Here, we collect some definitions and properties concerning p-process and supermartingales for the unfamiliar reader, and restate a technical lemma necessary for upcoming proofs.

Lemma 1 (Lemma 8 from JJ)**.** *Let $a \in \mathbb{R}_+^n$ be a n-dimensional vector with positive real entries, and for $i = 1, \ldots, n$ let Z_i be independent random variables where*

$$\mathbb{P}\left(Z_i \geq t\right) \leq \exp(-t/a_i).$$

Then for any $\delta \in (0, 1)$,

$$\sum_{i=1}^{n} Z_i \leq 5 \log(1/\delta) \sum_{i=1}^{n} a_i.$$

occurs with at least probability $1 - \delta$.

A.1 Equivalence property of p-processes

We note the following equivalence proposition for p-processes. Lemma 3 in Howard et al. [16] and Lemmas 1 and 2 in Ramdas et al. [31] makes similar statements regarding sequential processes, but do not additionally characterize the behavior of the infimum of a p-process.

Proposition 6. *The following statements are equivalent for a discrete-time process $(P_t)_{t\geq 1}$:*

(i) $(P_t)_{t\geq 1}$ *is a p-process i.e.* $\mathbb{P}(P_\tau \leq \alpha) \leq \alpha$ *for all (possibly infinite)* $\tau \in \mathcal{T}$ *and all* $\alpha \in (0,1)$*;*

(ii) $\mathbb{P}(P_\tau \leq \alpha) \leq \alpha$ *for all finite* $\tau \in \mathcal{T}$ *and all* $\alpha \in (0,1)$*;*

(iii) $\mathbb{P}(\exists t \geq 1 : P_t \leq \alpha) \leq \alpha$ *for all* $\alpha \in (0,1)$*;*

(iv) $\inf_{t\geq 1} P_t$ *is superuniformly distributed (its distribution is stochastically larger than uniform).*

Proof. In what follows, let $\tau_\alpha \in \mathcal{T}$ be defined as $\tau_\alpha := \inf\{t \geq 1 : P_t \leq \alpha\}$, which is defined to be infinite if P_t never drops below α.

(i)⇒(ii) is trivial by definition.

(ii)⇒(iii): Fix $\alpha \in (0,1)$. By (ii), we have $\mathbb{P}(P_{\tau_\alpha \wedge n} \leq \alpha) \leq \alpha$ for all $n \geq 1$. It follows that

$$\mathbb{P}(\exists t \geq 1 : P_t \leq \alpha) = \lim_{n\to\infty} \mathbb{P}(\exists t \in \{1,\dots,n\} : P_t \leq \alpha) = \lim_{n\to\infty} \mathbb{P}(P_{\tau_\alpha \wedge n} \leq \alpha) \leq \alpha.$$

(iii)⇒(iv): For each $\epsilon > 0$, since $\inf_{t\geq 1} P_t \leq \alpha$ implies $\tau_{\alpha+\epsilon} < \infty$, we have

$$\mathbb{P}\left(\inf_{t\geq 1} P_t \leq \alpha\right) \leq \mathbb{P}(\tau_{\alpha+\epsilon} < \infty) \leq \alpha + \epsilon.$$

As $\epsilon > 0$ is arbitrary, we get $\mathbb{P}(\inf_{t\geq 1} P_t \leq \alpha) \leq \alpha$, i.e., $\inf_{t\geq 1} P_t$ is superuniformly distributed.

(iv)⇒(i): For any $\tau \in \mathcal{T}$ and $\alpha \in (0,1)$, since $P_\tau \geq \inf_{t\geq 1} P_t$, we have $\mathbb{P}(P_\tau \leq \alpha) \leq \mathbb{P}(\inf_{t\geq 1} P_t \leq \alpha) \leq \alpha$, thus showing that (P_t) is a p-process. □

As a direct consequence of Proposition 6, if (P_t) is a p-process and P_s is uniformly distributed on $[0,1]$ for some $s \geq 1$, then we have $\mathbb{P}(P_s \leq P_t) = 1$ for all $t \geq 1$, since P_s is as small as $\min_{t\geq 1} P_t$. Therefore, if a p-process does not always take its minimum at a deterministic point s, then P_s cannot be uniformly distributed on $[0,1]$. In other words, for all deterministic t, the random variables P_t are, in general, not "precise" (i.e., uniform on $[0,1]$) p-variables, but conservative ones. In contrast, the random variables E_t from an e-process (E_t) are "precise" (i.e., have expectation 1) as soon as (E_t) is a nonnegative martingale starting at 1.

A.2 Nonnegative supermartingales

A real-valued process $(M_t)_{t\geq 0}$ is a supermartingale w.r.t. a filtration $(\mathcal{F}_t)$ if it satisfies:

$$\mathbb{E}[M_t | \mathcal{F}_{t-1}] \leq M_{t-1} \text{ for } t \in \mathbb{N}. \tag{6}$$

For nonnegative supermartingales, we typically assume $M_0 = 1$ for simplicity; they possess two useful properties. The first is the optional stopping theorem.

Fact 6 (Optional stopping theorem. Durrett [13], Ramdas et al. [31])**.** *Let (M_t) be a nonnegative supermartingale w.r.t. $(\mathcal{F}_t)$. Then, for any stopping time $\tau \in \mathcal{T}$:*

$$\mathbb{E}[M_\tau] \leq M_0.$$

The second is Ville's inequality.

Fact 7 (Ville's inequality)**.** *Let (M_t) be a nonnegative supermartingale w.r.t. $(\mathcal{F}_t)$. Let $s \in \mathbb{R}^+$ be a number in the positive reals.*

$$\mathbb{P}(\exists t \in \mathbb{N} : M_t \geq s) \leq \frac{M_0}{s}.$$

B Proofs

B.1 Proofs of results in Section 3.1

The proofs of Propositions 1 to 4 all follow from the application of one of Facts 1 to 3.

First, we note that $(P_{1,t}), \ldots, (P_{k,t})$ being p-processes implies that $P_{1,t}, \ldots, P_{k,t}$ are p-variables for all $t \in \mathbb{N}$. Thus, for any choice of stopping time $\tau^* \in \mathcal{T}$ for the algorithm, $P_{1,\tau^*}, \ldots, P_{k,\tau^*}$ are p-variables.

Consequently, Proposition 2 for the adaptive and dependent p-variables case and Proposition 3 for the adaptive and dependent p-variables with constrained rejection sets case follow from Fact 1 and Fact 2, respectively.

Similarly, we note that $E_{1,\tau^*}, \ldots E_{k,\tau^*}$ are e-variables, since $(E_{1,t}), \ldots, (E_{k,t})$ are e-processes. As a result, Proposition 4 follows from Fact 3.

Now, we prove Proposition 1 in a slightly different manner than JJ, using the notion of self-consistency.

Proof of Proposition 1. Consider an arbitrary $i \in [k]$. Recall that each $P_{i,t}$ is determined only by $X_{i,t_i(1)}, \ldots, X_{i,t_i(T(i))}$. By independence of $X_{i,t}$ across $i \in [k]$ and $t \in \mathbb{N}$, we rename $X_{i,t_i(j)}$ as $X_{i,j}$, since they are identically distributed. Thus, $P_{i,t}$ is now constructed from $X_{i,1}, \ldots, X_{i,t}$. We perform this transformation so we can consider $P_{i,t}$ in the infinite-sample limit. Under our renaming, $\mathcal{S}_{\tau^*}$ is the output of running BH on $P_{1,T_1(\tau^*)}, \ldots, P_{k,T_k(\tau^*)}$.

Define $P_i^* := \inf_{t \geq 1} P_{i,t}$. Note that P_i^* are independent p-variables across $i \in [k]$ by Proposition 6 since $X_{1,t}, \ldots, X_{k,t}$ are independent for all $t \in \mathbb{N}$ and $(P_{1,t}), \ldots, (P_{k,t})$ are p-processes. We can derive self-consistency w.r.t. P_i^* as follows.

$$\begin{aligned} \max_{i \in \mathcal{S}_{\tau^*}} P_i^* &\leq \max_{i \in \mathcal{S}_{\tau^*}} P_{i,T_i(\tau^*)} && \text{def. of } P_i^* \\ &\leq \frac{|\mathcal{S}_{\tau^*}|\delta'}{k}. && \text{p-self-consistency of } \mathcal{S}_{\tau^*} \end{aligned}$$

Combined with Fact 2, we can show that $\mathrm{FDR}(\mathcal{S}_{\tau^*}) \leq \delta' \log(1 + \log(1/\delta'))$.

Separately, we can also apply the FDR guarantee on the output of BH on arbitrarily dependent p-variables from Fact 1. Consequently, we can guarantee $\mathrm{FDR}(\mathcal{S}_{\tau^*}) \leq \delta' \log(1 + \log(1/\delta')) \wedge \delta' \log k$. Thus, our choice of δ' implies $\mathrm{FDR}(\mathcal{S}_{\tau^*}) \leq \delta$, which is our desired result. □

B.2 Proof of Proposition 5

Howard et al. [16] actually specifiy a more general form for λ_ℓ and w_ℓ for the discrete mixture e-process, $E_{i,t}^{\mathrm{DM}}$. Let f be a probability density over $(0, \lambda^{\max}]$ and nonincreasing over that interval, $\overline{\lambda} \in \mathbb{R}^+$ satisfy $\overline{\lambda} \leq \lambda^{\max}$, and $\eta > 1$ be a step size. Howard et al. [15] define λ_ℓ, w_ℓ as follows:

$$\lambda_\ell := \frac{\overline{\lambda}}{\eta^{\ell+1/2}} \text{ and } w_k := \frac{\overline{\lambda}(\eta - 1) f(\lambda_\ell \sqrt{\eta})}{\eta^{\ell+1}}. \tag{7}$$

Let,

$$f_s^{\mathrm{LIL}} := \frac{(s-1)s^{s-1}\mathbf{I}\{0 \leq \lambda \leq 1/e^s\}}{\lambda \log^s \lambda^{-1}},$$

for any $s > 1$. We will now connect these definitions to (3b). Set $\overline{\lambda} = 1/e$, $\eta = e$, and $f = f_2^{\mathrm{LIL}}$. Then,

$$\begin{aligned}
\lambda_\ell &= \frac{1}{e^{\ell+3/2}}, \\
w_\ell &= \frac{\frac{1}{e}(e-1) f_2^{\mathrm{LIL}}(\frac{1}{e^{\ell+3/2}} \cdot \sqrt{e})}{e^{\ell+1}} \\
&= \frac{(e-1) f_2^{\mathrm{LIL}}(\frac{1}{e^{\ell+1}})}{e^{\ell+2}} \\
&= \frac{(e-1)\frac{2\mathbf{I}\{\ell \geq 1\}}{\frac{1}{e^{\ell+1}} \log^2(e^{\ell+1})}}{e^{\ell+2}} \\
&= \frac{2(e-1)\mathbf{I}\{\ell \geq 1\}}{(\frac{1}{e^{\ell+1}})(e^{\ell+2}) \log^2(e^{\ell+1})} \\
&= \frac{2(e-1)\mathbf{I}\{\ell \geq 1\}}{e(\ell+1)^2}.
\end{aligned}$$

By reindexing ℓ, we can redefine the variables as follows:

$$\lambda_\ell = \frac{1}{e^{\ell+5/2}} \text{ and } w_\ell = \frac{2(e-1)}{e(\ell+2)^2}.$$

To prove Proposition 5, we prove the following more general proposition which is derived from existing results in Howard et al. [16].

Proposition 7 (Derived from equations (49) and (82) of Howard et al. [16])**.** *Let,*

$$E_{i,t} := \sum_{\ell=0}^{\infty} w_\ell \exp\left(\sum_{j=1}^{T_i(t)} \lambda_\ell (X_{i,t_i(j)} - \mu_0) - \frac{\lambda_\ell^2}{2}\right).$$

If $\sum_{\ell=0}^{\infty} w_\ell \leq 1$*, then* $(E_{i,t})$ *is a nonnegative supermartingale, and consequently an e-process, if the conditional distribution* $X_{i,t} \mid \mathcal{F}_{t-1}$ *is* 1*-sub-Gaussian and* $\mathbb{E}[X_{i,t} \mid \mathcal{F}_{t-1}] \leq \mu_0$ *for all* $t \in \mathbb{N}$*.*

Proof. Let

$$M_{i,t}^{\lambda} := \exp\left(\sum_{j=1}^{T_i(t)} \lambda (X_{i,t_i(j)} - \mu_0) - \frac{\lambda^2}{2}\right),$$

where $\lambda \in \mathbb{R}$. $(M_{i,t})$ is a nonnegative supermartingale because of the sub-Gaussian and bounded conditional mean assumptions on $X_{i,t}$. Let, $w_{\mathrm{sum}} = \sum_{\ell=0}^{\infty} w_\ell$. Now, we show that $E_{i,t}$ is a supermartingale:

$$\begin{aligned}
\mathbb{E}\left[E_{i,t} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{\ell=0}^{\infty} w_\ell M_{i,t}^{\lambda_\ell} \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{\ell=0}^{\infty} w_\ell \mathbb{E}\left[M_{i,t}^{\lambda_\ell} \mid \mathcal{F}_{t-1}\right] \\
&\leq \sum_{\ell=0}^{\infty} w_\ell M_{i,t-1}^{\lambda_\ell} \\
&= E_{i,t-1}.
\end{aligned}$$

The sole inequality is by the supermartingale property of $(M_{i,t}^{\lambda_\ell})$. Thus, we have shown our desired result. □

B.3 Proof of Theorem 1

We follow a similar path as the sample complexity proof (i.e. Theorem 2) from JJ for Theorem 1. Our goal is to show that we reject the following set with at least $1-\delta$ probability:

$$\mathcal{R} = \{i \in \mathcal{H}_1 : \widehat{\mu}_{i,t} + \varphi(t,\delta) \geq \mu_i \text{ for all } t \in \mathbb{N}\}. \tag{8}$$

Lemma 2. $\mathbb{E}[|\mathcal{R}|] \geq (1-\delta)|\mathcal{H}_1|$.

Proof. We have the following:

$$\begin{aligned}
\mathbb{E}[|\mathcal{R}|] &= \sum_{i \in \mathcal{H}_1} \mathbb{P}\left(\widehat{\mu}_{i,t} + \varphi(t,\delta) \geq \mu_i\right) \\
&\geq \sum_{i \in \mathcal{H}_1} \mathbb{P}\left(|\widehat{\mu}_{i,t} - \mu_i| \leq \varphi(t,\delta)\right) \\
&\geq (1-\delta)|\mathcal{H}_1|. && \text{Fact 4}
\end{aligned}$$

□

Lemma 2 shows that rejecting $\mathcal{R}$ is sufficient to produce rejection sets that have $\text{TPR}(\mathcal{S}) \geq 1-\delta$. Thus, our goal in this proof is to show a bound on $T \coloneqq \min\{t \in \mathbb{N} : \mathcal{R} \subseteq \mathcal{S}_t\}$ with at least $1-\delta$ probability, where $T = \infty$ if $\mathcal{R} \not\subseteq \mathcal{S}_t$ for all $t \in \mathbb{N}$. Note that for all $t \geq T$, $\mathcal{S}_T \subseteq \mathcal{S}_t$ by the way (5a) is defined — it does not sample arms that have already been rejected.

We note that we can use any φ defined in Fact 4 in this proof and still achieve the desired result. For simplicity, we use φ to denote φ^0 in this proof. First we define a notion of inverse for φ. Let

$$\varphi^{-1}(\epsilon,\delta) \coloneqq \min\{t : \varphi(t,\delta) \leq \epsilon\}. \tag{9}$$

JJ and other work [18] show that for some absolute constant $c > 0$,

$$\varphi^{-1}(\epsilon,\delta) \leq c\epsilon^{-2}\log(\log(\epsilon^{-2})/\delta) \text{ for all } \epsilon \in \mathbb{R}^+, \delta \in (0,1). \tag{10}$$

Also, recall that $f \lesssim g$ denotes f asymptotically dominates g i.e. there exist $c > 0$ that is independent of the problem parameters such that $f \leq cg$.

We decompose T into the number of time steps the algorithm samples a null arm, and the number of time steps the algorithm samples a non-null arm:

$$T = \sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} = \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} + \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\}. \tag{11}$$

Our first goal is to prove a sample complexity bound on $\sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\}$. We define the following variables for each $i \in [k]$.

$$\rho_i \coloneqq \inf\{\rho \in [0,1] : |\widehat{\mu}_{i,t} - \mu_i| > \varphi(t,\rho) \text{ for all } t \in \mathbb{N}\} \cup \{1\}. \tag{12}$$

Lemma 3. *For each $i \in [k]$, $\mathbb{P}(\rho_i \leq s) \leq s$ for $s \in (0,1)$ i.e. ρ_i is superuniformly distributed.*

The above lemma follows directly from Fact 4. We also define a concentration bound for independent superuniformly distributed variables.

Lemma 4. *For any fixed positive reals $a_1, \ldots, a_d$, independent superuniformly distributed random variables $r_1, \ldots, r_d$, and $\beta \in (0,1)$, the following event occurs with probability at least $1-\beta$:*

$$\sum_{i=1}^{d} a_i \log(1/r_i) \leq 5\log(1/\beta)\sum_{i=1}^{d} a_i.$$

This lemma follows directly from recognizing $a_i \log(1/r_i)$ satisfies the requirements for Z_i in Lemma 1. Now, we will show that the UCB for each $i \in \mathcal{R}$ will be above μ_i.

Lemma 5. *Let ν_i be sub-Gaussian for each $i \in [k]$. Any algorithm with $(\mathcal{A}_t)$ that outputs $\mathcal{I}_t = \{I_t\}$ as defined in (5a) has the following property:*

$$\sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta\rho_i).$$

Proof. The following is true for any $i \in \mathcal{R}$ and $t \in \mathbb{N}$:

$$\begin{aligned} \widehat{\mu}_{i,t} + \varphi(T_i(t), \delta)) &\geq \mu_i - \varphi(T_i(t), \rho_i) + \varphi(T_i(t), \delta)) && \text{by def. of } \rho_i \text{ and } \mathcal{R} \\ &\geq \mu_i. && \text{by def. of } \mathcal{R} \end{aligned}$$

Thus, $\{\mathcal{R} \not\subseteq \mathcal{S}_t\}$ implies for any $t \in \mathbb{N}$:

$$\begin{aligned} \operatorname{argmax}_{i \in [k] \setminus \mathcal{S}_t} \widehat{\mu}_{i,t} + \varphi(T_i(t), \delta) &\overset{\text{(i)}}{\geq} \min_{i \in \mathcal{R}} \mu_i \\ &\geq \min_{i \in \mathcal{H}_1} \mu_i, \end{aligned} \tag{13}$$

where inequality (i) is by the definition of $\mathcal{R}$.

In addition, we argue that the UCB for $i \in \mathcal{H}_0$ will shrink below $\min_{i \in \mathcal{H}_1} \mu_i$ quickly. For $i \in \mathcal{H}_0$, the following is true for any $t \in \mathbb{N}$:

$$\begin{aligned} \widehat{\mu}_{i,t} + \varphi(T_i(t), \delta) &\leq \mu_i + \varphi(T_i(t), \rho_i) + \varphi(T_i(t), \delta) \\ &\leq \mu_i + 2\varphi(T_i(t), \delta\rho_i). \end{aligned} \tag{14}$$

Thus, $\{\forall i \in \mathcal{H}_0 : \mu_i + 2\varphi(T_i(t), \delta\rho_i) \leq \min_{i \in \mathcal{H}_1} \mu_i,\ \mathcal{R} \not\subseteq \mathcal{S}_t\} \implies \{I_t \in \mathcal{H}_1\}$ for all $t \in \mathbb{N}$ by (13) and (14).

Subsequently, we argue the following:

$$\begin{aligned} \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} &\leq \sum_{t=1}^{\infty} \mathbf{I}\left\{\exists i \in \mathcal{H}_0 : \mu_i + 2\varphi(T_i(t), \delta\rho_i) > \min_{i \in \mathcal{H}_1} \mu_i, \mathcal{R} \not\subseteq \mathcal{S}_t\right\} \\ &\leq \sum_{i \in \mathcal{H}_0} \varphi^{-1}(\Delta_i/2, \delta\rho_i) && \mu_i \leq \mu_0 \text{ for all } i \in \mathcal{H}_0 \\ &\lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta\rho_i). \end{aligned}$$

Thus, we have shown our desired result. □

Now, we proceed to show a bound on $\sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\}$. Denote π as an arbitrary mapping from $\mathcal{H}_1$ to $[|\mathcal{H}_1|]$. Let $(x)_+ = x \vee 0$ for any $x \in \mathbb{R}$. We define additional variables as follows:

$$\begin{aligned} \ell_i' &:= (\lceil \log(2\Delta_i^{-1}) - 5/2 \rceil)_+, \\ \rho_i^{\mathrm{DM}} &:= \min_{t \in \mathbb{N}} \frac{1}{\exp\left(\sum_{j=1}^{T_i(t)} \lambda_{\ell_i'}(\mu_i - X_{i,t_i(j)}) - \lambda_{\ell_i'}^2/2\right)}. \end{aligned}$$

Lemma 6. $\mathbb{P}\left(\rho_i^{\mathrm{DM}} \leq s\right) \leq s$ *for* $s \in (0,1)$ *i.e.* ρ_i^{DM} *is superuniformly distributed for each* $i \in \mathcal{H}_1$.

Proof. First, we prove an underlying process is a nonnegative supermartingale. Let

$$M_t = \exp\left(\sum_{j=1}^{T_i(t)} \lambda_{\ell'_i}(\mu_i - X_{i,t_i(j)}) - \lambda_{\ell'_i}^2/2\right).$$

Assume that arm i is selected at time t — otherwise the supermartingale property is directly satisfied.

$$\begin{aligned}
\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\sum_{j=1}^{T_i(t)} \lambda_{\ell'_i}(\mu_i - X_{i,t_i(j)}) - \lambda_{\ell'_i}^2/2\right) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\exp\left(\lambda_{\ell'_i}(\mu_i - X_{i,t}) - \lambda_{\ell'_i}^2/2\right) \mid \mathcal{F}_{t-1}\right] \exp\left(\sum_{j=1}^{T_i(t-1)} \lambda_{\ell'_i}(\mu_i - X_{i,t_i(j)}) - \lambda_{\ell'_i}^2/2\right) \\
&\leq M_{t-1},
\end{aligned}$$

where the final inequality holds because $X_{i,t}$ are i.i.d. across $t \in \mathbb{N}$, have mean μ_i, and are 1-sub-Gaussian.

Thus, ρ_i^{DM} is a superuniform random variable by applying Ville's inequality to (M_t).

□

Proposition 8 (Growth of $E_{i,t}^{\mathrm{DM}}$)**.** *When $i \in \mathcal{H}_1$ and ν_i is 1-sub-Gaussian,*

$$\log E_{i,t} \gtrsim \Delta_i^2 T_i(t) - \log\log(\Delta_i^{-2}) - \log(1/\rho_i^{\mathrm{DM}}).$$

Proof. We show the following lower bound on $E_{i,t}^{\mathrm{DM}}$:

$$\begin{aligned}
E_{i,t}^{\mathrm{DM}} &= \sum_{\ell=0}^{\infty} w_\ell \exp(T_i(t)(\lambda_\ell \Delta_i - \lambda_\ell^2)) \exp\left(\sum_{j=1}^{T_i(t)} \lambda_\ell^2/2 - \lambda_\ell(\mu_i - X_{i,t_i(j)})\right) \\
&\geq w_{\ell'_i} \exp(T_i(t)(\lambda_{\ell'_i} \Delta_i - \lambda_{\ell'_i}^2)) \exp\left(\sum_{j=1}^{T_i(t)} \lambda_{\ell'_i}^2/2 - \lambda_{\ell'_i}(\mu_i - X_{i,t_i(j)})\right) \\
&\geq \exp\left(\frac{1}{4e}\Delta_i^2 T_i(t) - \log(1/w_{\ell'_i}) - \log(1/\rho_i^{\mathrm{DM}})\right). && \text{by def. of } \ell'_i \text{ and } \rho_i^{\mathrm{DM}}
\end{aligned}$$

Thus, plugging in $w_{\ell'_i}$, we get our desired result. □

Proof of Theorem 1. By Proposition 8,

$$T_i(t) \gtrsim \Delta_i^{-2} \log(\varepsilon \log(\Delta_i^{-2})/\rho_i^{\mathrm{DM}})$$

implies $E_{i,t} \geq \varepsilon$ for $\varepsilon > 0$.

Now, we can derive the following bound:

$$\begin{aligned}
\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} &= \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} + \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\} \\
&\lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log\log(\Delta_i^{-2}) + \Delta_i^{-2} \log(1/\rho_i) + \Delta_i^{-2} \log(1/\delta) \\
&\quad + \max_{\pi} \sum_{i \in \mathcal{H}_1} \Delta_i^{-2} \log\log(\Delta_i^{-2}) + \Delta_i^{-2} \log(1/\rho_i^{\mathrm{DM}}) + \Delta_i^{-2} \log(k/\pi(i)\delta). && \text{by Lemma 5}
\end{aligned}$$

Algorithm 2: An generic algorithm that uses a BAI subroutine to find a best arm that is not in the rejection set for the algorithm to then repeatedly sample and eventually reject.

Input: A BAI algorithm $\mathcal{B}$ that takes in $B \subseteq [k]$, and a history of samples and initial randomness $D_t(B) \coloneqq U \cup \{(i, j, X_{i,j}) : j \leq t, i \in \mathcal{I}_j \cap B\}$. At each step, $\mathcal{B}$ outputs a superarm $\mathcal{I} \in \mathcal{K}$ to sample next, or a best arm $i \in B$. Let $\delta \in (0, 1)$ be the level of FDR control and $\delta' \in (0, 1)$ be the corrected level for p-variables. Let $(e_{i,t})$ and $(p_{i,t})$ be realized values of e-processes and p-processes, respectively, for each $i \in [k]$. Let $\tau^* \in \mathcal{T}$ be the stopping time for the algorithm.

Initialize $\mathcal{S}_0 \coloneqq \emptyset$

Initialize bestarm $\coloneqq$ **none**

for $t \in 1, \ldots,$ **do**

  $B \coloneqq [k] \setminus \mathcal{S}_{t-1}$

  if bestarm *is* ***none*** *or* bestarm $\in \mathcal{S}_{t-1}$ **then**

    $\mathcal{I}_t \coloneqq \mathcal{B}(B, D_{t-1}(B))$

    if $\mathcal{B}(B, D_{t-1}(B))$ *terminated with best arm* I_t **then** bestarm $\coloneqq I_t$;

  else

    $\mathcal{I}_t \coloneqq \{\text{bestarm}\}$ (or an arbitrary $\mathcal{I} \in \mathcal{K}$ such that bestarm $\in \mathcal{K}$).

  end

  Update e-process or p-process for each queried arm not in $\mathcal{S}_{t-1}$.

  $$\mathcal{S}_t \coloneqq \begin{cases} \mathrm{BH}[\delta'](p_{1,t}, \ldots, p_{k,t}) \text{ or arbitrary p-self-consistent set} & \text{if using p-variables} \\ \mathrm{eBH}[\delta](e_{1,t}, \ldots, e_{k,t}) \text{ or arbitrary e-self-consistent set} & \text{if using e-variables} \end{cases}$$

  if $\tau^* = t$ **then** stop and **return** $\mathcal{S}_t$;

end

Recall that ρ_i, by Lemma 3, and ρ_i^{DM}, by Lemma 6, are superuniform random variables that are independent across $i \in [k]$ and $i \in \mathcal{H}_1$, respectively. Consequently, we can apply Lemma 4 at level $\beta = \delta/2$ to ρ_i for $i \in [k]$ and ρ_i^{DM} for $i \in \mathcal{H}_1$. Then, the following happens with at least $1 - \delta$ probability:

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \max_{\pi} \sum_{i \in \mathcal{H}_1} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \Delta_i^{-2} \log(k/\pi(i)).$$

We can derive two different bounds. The first is using the fact that $\sum_{i=1}^{|\mathcal{H}_1|} \log(k/i) \leq k$. As a result,

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim k\Delta^{-2} \log(\log(\Delta^{-2})/\delta).$$

The second comes from dropping the $\pi(i)$ term, which is as follows:

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \sum_{i \in \mathcal{H}_1} \Delta_i^{-2} \log(k \log(\Delta_i^{-2})/\delta).$$

Thus, we have shown both sample complexity bounds as desired. □

C Generic algorithms for $(\mathcal{A}_t)$

We propose two generic algorithms that can be used for the exploration component in Algorithm 1 regardless of the type of hypotheses tested or what the joint distribution of $X_{1,t}, \ldots, X_{k,t}$ is. For simplicity, we assume that the algorithm can always sample each arm separately, i.e. $\{\{1\}, \ldots, \{k\}\} \subseteq \mathcal{K}$.

Algorithm 3: An generic algorithm applicable to any combinatorial bandit and set of hypotheses that utilizes the evidence itself (i.e. p-variables or e-variables) to select arms to sample.

Input: Let $\delta \in (0,1)$ be the level of FDR control and $\delta' \in (0,1)$ be the corrected level for p-variables. $(e_{i,t})$ and $(p_{i,t})$ be realized values of e-processes and p-processes, respectively, for each $i \in [k]$. Let $\tau^* \in \mathcal{T}$ be the stopping time for the algorithm.

Initialize $\mathcal{S}_0 := \emptyset$

Initialize $e_{i,0} = 1$ or $p_{i,0} = 1$ for all $i \in [k]$

for $t \in 1, \ldots,$ **do**

  if $t \leq k$ **then**

    $I_t := t$

    $\mathcal{I}_t := \{I_t\}$ or an arbitrary $\mathcal{I} \in \mathcal{K}$ where $I_t \in \mathcal{I}$

  else

    $I_t \in \begin{cases} \operatorname{argmin}_{i \in [k] \setminus \mathcal{S}_{t-1}} p_{i,t-1} & \text{if using p-variables} \\ \operatorname{argmax}_{i \in [k] \setminus \mathcal{S}_{t-1}} e_{i,t-1} & \text{if using e-variables} \end{cases}$ (an arbitrary element of argmin/argmax).

    $\mathcal{I}_t := \{I_t\}$ or an arbitrary $\mathcal{I} \in \mathcal{K}$ where $I_t \in \mathcal{I}$

  end

  Update e-process or p-process for each queried arm not in $\mathcal{S}_{t-1}$.

  $\mathcal{S}_t := \begin{cases} \text{BH}[\delta'](p_{1,t}, \ldots, p_{k,t}) \text{ or arbitrary p-self-consistent set} & \text{if using p-variables} \\ \text{eBH}[\delta](e_{1,t}, \ldots, e_{k,t}) \text{ or arbitrary e-self-consistent set} & \text{if using e-variables} \end{cases}$

  if $\tau^* = t$ **then** stop and **return** $\mathcal{S}_t$;

end

Reduction to best arm identification (BAI) The first relies on having access to a best arm identification (BAI) algorithm. BAI is well studied problem, and there exist many algorithms for it in both the standard bandit setting [1, 22, 18, 23] and combinatorial bandit settings [12, 10, 20]. A BAI algorithm returns the "best arm" i.e. the arm with the highest mean reward, with high probability. Thus, we can employ a BAI algorithm as a subroutine to repeatedly find the best arm out of arms not in the rejection set, and then repeatedly sample that best arm until it is rejected. Algorithm 2 formulates an algorithm using a BAI subroutine that fits the meta-algorithm introduced in Algorithm 1 . Consequently, we can immediately have access to algorithms for multiple testing that have non-trivial exploration components for a wide variety of settings.

Largest e-process (or smallest p-process) If no apparent exploration strategy exists, we can always select the arm that currently has the most evidence for rejection, but has not yet been rejected. Algorithm 3 illustrates this algorithm — our exploration strategy is to simply pick the superarms that contains the arm that already has the "most" evidence (largest e-value or smallest p-value). Thus, simply having e-variables or p-variables for the hypotheses we are testing can be used to inform the sampling strategy.

Both of the aforementioned algorithms guarantee FDR control due to being instances of Algorithm 1.

Proposition 9. *Algorithms 2 and 3 guarantee that* $\sup_{\tau \in \mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \leq \delta$.

As a result, we always have a default choice of exploration component if we are unaware of any domain specific strategies for sampling.

D Extensions on the bandit setting

In this section, we consider some special cases and extensions on the bandit settings. This includes settings involving streaming data, constrained rejection sets, multiple agents, and hypotheses involving multiples arms. Critically, we show how our framework can be easily adaptable to each of these settings to still maintain valid FDR guarantees.

D.1 Streaming data setting

A unique instance of the combinatorial bandits is the streaming data setting, where the algorithm has access to the all rewards at each time step. Instead of choosing a sampling policy, the algorithm can choose a stopping time τ_i for each arm $i \in [k]$ that marks when the algorithm will cease observing arm i. Although $X_{1,t}, \ldots, X_{k,t}$ may be arbitrarily dependent, Algorithm 1 with e-variables can still use all the observations from each arm at each time step. This is because e-BH on e-variables maintains the same FDR control irrespective of dependence structure. Thus, we can propose a simple strategy in Algorithm 4 that stops the monitoring of an arm once that arm has been rejected by e-BH, and can limit the amount of time between rejections or total time run before the algorithm stops. By Proposition 4, we have the following FDR guarantee.

Proposition 10. *Algorithm 4 ensures* $\sup_{\tau \in \mathcal{T}} \mathrm{FDR}(\mathcal{S}^*_\tau) \leq \delta$.

Bartroff and Song [5] also study multiple testing in the streaming data setting, and prove FDR guarantees similar to Proposition 10 for an algorithm that is virtually identical to Algorithm 4 with p-variables and BH. A key difference between their results and ours is that they use test statistics in their algorithm instead of p-variables, and make assumptions about the power of the test statistics that also allow them to provide guarantees about the false negative rate i.e. the expected proportion of hypotheses that are not rejected which are true discoveries. Thus, our framework for FDR control subsumes existing methods for the streaming setting. Other error metrics such as family-wise error rate and probabilistic bounds on the FDP have also been studied in the sequential setting [3, 4, 2].

Algorithm 4: An algorithm for monitoring in the streaming data setting. This algorithm stops when the maximum time $t_{\max}$ has been reached, or more than t_{gap} steps have passed since the last rejection. Once an arm is added to $\mathcal{S}_t$, the algorithm stops monitoring it.

$\mathcal{S}_0 = \emptyset$
$t_{\text{prev.rejection}} = 0$
for $t \in 1, \ldots, t_{\max}$ **do**
  $\mathcal{I}_t \coloneqq [k] \setminus \mathcal{S}_{t-1}$
  $\mathcal{S}_t \coloneqq \begin{cases} \mathrm{eBH}[\delta](e_{1,t}, \ldots, e_{k,t}) & \text{if using e-variables} \\ \mathrm{BH}[\delta'](p_{1,t}, \ldots, p_{k,t}) & \text{if using p-variables} \end{cases}$
  if $t - t_{\text{prev.rejection}} > t_{\text{gap}}$ ***or*** $\mathcal{S}_t = [k]$ **then return** $\mathcal{S}_t$;
end
return $\mathcal{S}_t$

D.2 Structured rejection sets

Structured rejection sets arise in problems where there is a fixed hierarchy that restrict the sets of hypotheses that can be rejected e.g. hypothesis 2 can only be rejected if hypothesis 1 is rejected also. Recent work in multiple testing with FDR control has studied settings with general structural constraints [25] and when the constraints have been restricted to form a directed acyclic graph (DAG) [30]. A DAG constraint requires all predecessors of a hypothesis in the DAG to be rejected before the hypothesis itself can be rejected. Thus, the algorithm does not necessarily output the result of BH or e-BH, but rather a p-self-consistent or e-self-consistent set, respectively. Table 2 illustrates the FDR guarantees for p-variables in the structured setting for different dependence relationsips between $X_{1,t}, \ldots, X_{k,t}$. In case with adaptive $(\mathcal{A}_t)$ and τ^* when $X_{1,t} \ldots, X_{k,t}$ are independent — the guarantee in that setting remains unchanged due to the proof of FDR control already being based upon the fact that the output of BH was p-self-consistent to $P^*_1, \ldots, P^*_k$. Similarly, e-variables still do not pay a penalty when moving from e-BH to an arbitrary e-self-consistent set. The FDR when using e-variables remains below δ after setting $\alpha = \delta$.

We show an example in Figure 3 of a set of hypotheses in a DAG structure. Thus, a hypothesis can only be rejected if its predecessors in the DAG are also rejected. We compare the output of BH, e-BH, and both the largest e-self-consistent set and p-self-consistent set that respect the DAG constraints. The e-values are

Table 2: FDR control guaranteed by an arbitrary p-self-consistent set, and the δ' to ensure δ control of FDR in Algorithm 1 under different dependence structures and adaptivity of $(\mathcal{A}_t)$. Adaptive and non-adaptive strategies no longer have different guarantees when outputting a p-self-consistent set. On the other hand, the FDR control of an e-self-consistent set remains unchanged at $\alpha = \delta$.

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

calculated assuming that the p-variables are reciprocals of e-variables. We assume the p-variables are all arbitrarily dependent. The largest e-self-consistent set and the largest p-self-consistent set are simply the largest subset of e-BH and BH, respectively, that satisfies the DAG constraints.

D.3 Multiple agents

There are many scenarios where multiple agents are interacting with the same bandit and we hope to have the agents cooperatively accumulate evidence. For example, a research group could be interested in resuming the study of a hypothesis that previous researchers have run experiments on, and would like to combine existing evidence with the new evidence they collect from their own experiments. A cooperative situation could also arise when there are multiple groups that each work on a subset of some overarching set of hypotheses — the groups can combine the evidence they have for each hypothesis. In these cases, the evidence shared, either from previous studies or concurrent collaborators, might only be in the form of an e-value or p-value — the actual samples may be obfuscated for privacy reasons. Thus, each of the scenarios require the merging of multiple statistics (from each agent) into a single statistic representing the total amount of evidence for rejecting a hypothesis.

Assume we have m agents and let $E_1, \dots, E_m$ denote the e-variables all testing the same hypothesis. If the e-variables are all independent, we can define an ie-merging function (outputs an e-variable from independent e-variables) f_{prod} as follows:

$$f_{\mathrm{prod}}(E_1, \dots, E_m) \coloneqq \prod_{i=1}^{m} E_i.$$

Proposition 11. *If $E_1, \dots, E_m$ are independent e-variables, then $f_{\mathrm{prod}}(E_1, \dots, E_m)$ is also an e-variable.*

The above proposition follows from the fact that the expectation of the product is the product of expectation for independent random variables.

If $E_1, \dots E_m$ are dependent, then we can define the following e-merging function (outputs an e-variable from *arbitrarily dependent* e-variables):

$$f_{\mathrm{mean}}(E_1, \dots, E_m) \coloneqq \frac{1}{m} \sum_{i=1}^{m} E_m.$$

Proposition 12. *If $E_1, \dots, E_m$ are arbitrarily dependent e-variables, then $f_{\mathrm{mean}}(E_1, \dots, E_m)$ is also an e-variable.*

Vovk and Wang [41] show that the set of functions corresponding to all convex combinations of f_{mean} and 1 are the only admissible e-merging functions in the class of all symmetric e-merging functions. They also show a weaker sense of dominance for f_{prod} — they prove it outputs a larger e-value than any other symmetric ie-merging function if all the input e-values are at least 1. Thus, e-variables can be merged in a relatively simple fashion without many assumptions.

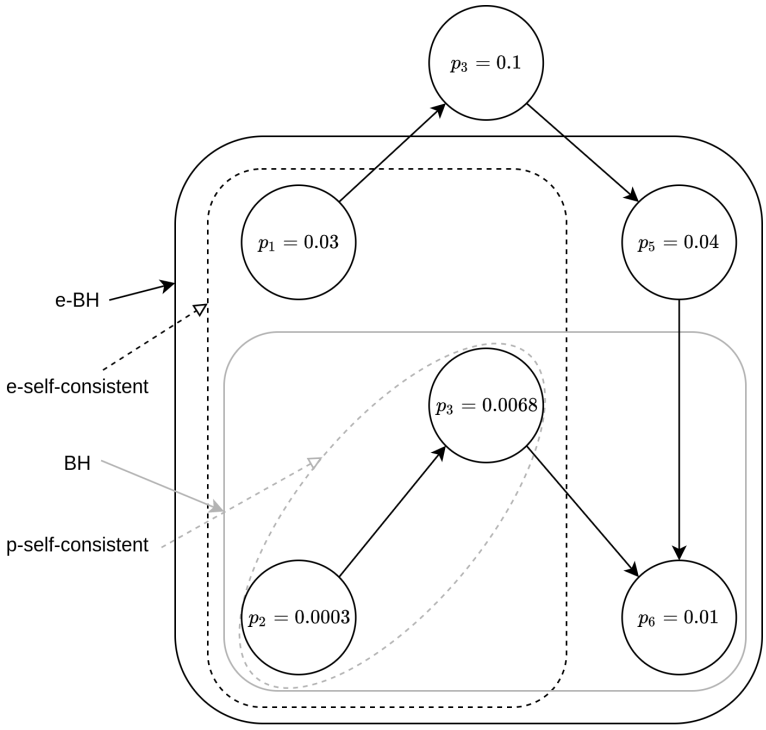


Figure 3: Example set of p-values for hypotheses that have a DAG constraint upon them and rejection sets that ensure $\text{FDR}(\mathcal{S}) \leq \delta = 0.05$. We assume the p-variables are arbitrarily dependent, and are reciprocals of e-variables for the sake of comparing e-variable vs. p-variable procedures. The e-self-consistent and p-self-consistent rejection sets are the largest such sets that satisfy the FDR guarantee and the DAG constraints. The e-BH and BH rejection sets violate the DAG constraints, i.e. they are not valid rejection sets, but they do maintain $\text{FDR}(\mathcal{S}) \leq \delta$. The largest valid e-self-consistent rejection set and p-self-consistent rejection set are simply the largest subsets that satisfy the DAG constraints of e-BH and BH, respectively.

On the other hand, merging p-variables is difficult. When the p-variables are independent, Birnbaum [8] show that any valid merging function which is monotonic w.r.t. to the p-values is admissible. When the p-variables are arbitrarily dependent, Vovk et al. [43] prove that there are also many admissible symmetric p-merging functions. Consequently, the p-variable picture is much less clear about how to optimally merge p-variables, particularly when there is arbitrary dependence among them. To illustrate these e-merging/ie-merging functions can be used in a bandit setting, we consider an example multi-agent problem where many research groups are submitting studies to the same journal.

D.3.1 Example: controlling the FDR of results in a journal

We consider a situation where the editors of a journal are interested in guaranteeing the accuracy of the results published within the journal. Specifically, they aim to ensure FDR control on the discoveries within the papers accepted to the journal. The journal requires that each study that is submitted is also accompanied by an e-value. Since many groups can be testing the same hypothesis, the journal can use the aforementioned merging techniques to combine the e-values reported by different groups and produce a valid, aggregate e-variable.

Formalizing the multi-agent setup The reward of the ith arm on the tth day for the ℓth agent is denoted as $X_{i,t}^{(\ell)}$ for all $t, \ell \in \mathbb{N}$ and $i \in [k]$. We let the index for agents, ℓ, be in $\mathbb{N}$ to allow for arbitrarily large, but finite, number of agents at each time step. We let the joint distribution of $X_{1,t}^{(\ell)}, \ldots, X_{k,t}^{(\ell)}$ be identically distributed across $\ell, t \in \mathbb{N}$. Consequently, the rewards $X_{1,t}^{(\ell)}, \ldots, X_{k,t}^{(\ell)}$ corresponding to each agent ℓ are identical in a marginal sense across all $\ell \in \mathbb{N}$. However, there can be arbitrary dependencies between the

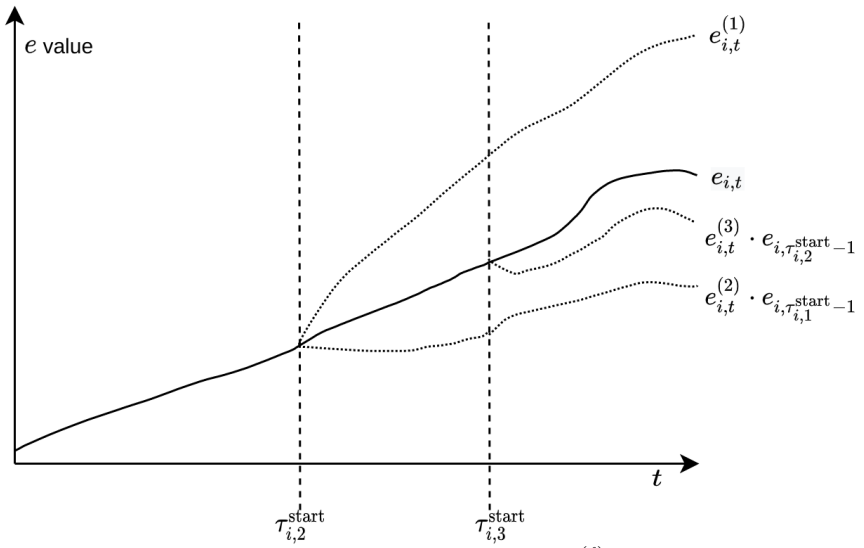


Figure 4: An illustration of how $e_{i,t}$ changes in relation to each $e_{i,t}^{(\ell)}$ for a case where $\ell = 3$ in Algorithm 5. We see that $e_{i,t}$ and $e_{i,t}^{(1)}$ are identical up to $\tau_{i,2}^{\text{start}} - 1$, where agent 2 begins to sample arm i. Agent 2's process starts at $e_{i,\tau_{i,2}^{\text{start}}-1}$. Similarly, Agent 3's process starts at $e_{i,\tau_{i,3}^{\text{start}}-1}$ when it starts to sample i as well. Validity of Algorithm 5 arises from the fact that each new agent ℓ has its own $e_{i,t}^{(\ell)}$ scaled by the $e_{i,t}$ that has been achieved already.

rewards of different agents. Thus, we allow for a setting where, for each $i \in [k]$ and $t \in \mathbb{N}$, $X_{i,t}^{(\ell)}$ is the same reward across all $\ell \in \mathbb{N}$, and a setting where $X_{i,t}^{(\ell)}$ are independent across $\ell \in \mathbb{N}$. Each agent $\ell \in \mathbb{N}$ outputs $\mathcal{I}_t^{(\ell)} \in \mathcal{K} \cup \{\emptyset\}$ for each $t \in \mathbb{N}$. Let the set of agents (e.g. set of studies) on day t that are testing hypothesis i be $A_{i,t}$ for each $t \in \mathbb{N}$ and $i \in [k]$. Critically, we require that $A_{i,t}$ be of finite cardinality almost surely and predictable w.r.t. the new canonical filtration $(\mathcal{G}_t)$. We define the canonical filtration for the multi-agent setting as follows:

$$\mathcal{G}_t \coloneqq \sigma(U \cup \{(i, s, \ell, X_{i,s}^{(\ell)}) : i \in \mathcal{I}_s^{(\ell)}, s \leq t, \ell \in A_{i,s}\}).$$

We denote the e-process of the ℓth agent for hypothesis i to be $(E_{i,t}^{(\ell)})$, where $E_{i,t}^{(\ell)} = 1$ if $\ell \notin A_{i,t}$. Implicitly, there exists a stopping time $\tau_{i,\ell}^{\text{start}}$ w.r.t. $(\mathcal{G}_t)$ that denotes the time when the ℓth agent begins testing the ith hypothesis for $\ell \in \mathbb{N}$ and $i \in [k]$ (i.e. the time of the first sample of arm i by agent ℓ). Algorithm 5 explicitly formulates the algorithm for dealing with e-values coming from multiple agents.

Figure 4 illustrates how $e_{i,t}$ behaves w.r.t. to the $e_{i,t}^{(\ell)}$ of each agent in Algorithm 5. Algorithm 5 uses a merging approach that is in between f_{prod} and f_{mean}. Intuitively, we know the rewards across $t \in \mathbb{N}$ are independent, and consequently we can merge e-values by taking the product. When merging across different $\ell \in \mathbb{N}, i \in [k]$, however, there may be arbitrary dependence between rewards. Consequently, we must take the mean of those e-values. From a betting perspective as discussed in Shafer [33], we can view our algorithm as splitting the current wealth (current $e_{i,t}$) evenly across each agent whenever a new agent is introduced before allowing each agent to continue or begin its own strategy. Regardless, we can show the following guarantee concerning Algorithm 5.

Proposition 13. *Let $(E_{i,t}^{(\ell)})$ be upper bounded by some nonnegative supermartingale $(M_{i,t}^{(\ell)})$ w.r.t. $(\mathcal{G}_t)$ for $i \in [k], \ell \in \mathbb{N}$ where $E_{i,t}^{(\ell)} = M_{i,t}^{(\ell)} = 1$ for $t < \tau_{i,\ell}^{\text{start}}$. Algorithm 5 ensures that $\sup_{\tau \in \mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \leq \delta$.*

Proof. Define $M_{i,t} \coloneqq \frac{1}{|A_{i,t}|} \sum_{\ell \in A_{i,t}} M_{i,t}^{(\ell)} \cdot M_{i,\tau_{i,\ell}^{\text{start}}-1}$ when $i \notin \mathcal{S}_{t-1}$ and $M_{i,t} \coloneqq M_{i,t-1}$ otherwise. We can see that $M_{i,t}$ upper bounds $E_{i,t}$ for all $t \in \mathbb{N}, i \in [k]$. We will show that $(M_{i,t})$ is a nonnegative supermartingale.

Algorithm 5: An algorithm for aggregating evidence in the form of e-values from many agents. The algorithm takes the mean of the e-values for each hypothesis on each day and applies an e-self-consistent procedure to these aggregated e-values to maintain valid FDR control at δ.

Input: A level of control δ in $(0, 1)$. $(e_{i,t}^{(\ell)})$ are the realized values of an e-process for $\ell \in \mathbb{N}, i \in [k]$.
Initialize $e_{i,0} \coloneqq 1$ for $i \in [k]$
$\mathcal{S}_0 \coloneqq \emptyset$
for $t \in 1, \dots$ **do**
  Receive new results from new or existing agents and update $e_{i,t}^{(j)}$ for $i \in [k]$ and $j \in [a_t]$
  for $i \in [k]$ **do**
    $$e_{i,t} \coloneqq \begin{cases} \frac{1}{|A_{i,t}|} \sum_{j \in A_{i,t}} e_{i,\tau_{i,j}^{\text{start}}-1} \cdot e_{i,t}^{(\ell)} & \text{if } i \notin \mathcal{S}_{t-1} \\ e_{i,t-1} & \text{else} \end{cases}.$$
  end
  $\mathcal{S}_t \coloneqq \text{eBH}[\delta](e_{1,t}, \dots, e_{k,t})$ or an arbitrary e-self-consistent set.
end

Assume that we have not rejected the ith hypothesis yet, since otherwise $M_{i,t} = M_{i,t-1}$, which satisfies the supermartingale property.

$$\begin{aligned}
\mathbb{E}[M_{i,t} \mid \mathcal{G}_{t-1}] &= \mathbb{E}\left[\frac{1}{|A_{i,t}|} \sum_{\ell \in A_{i,t}} M_{i,t}^{(\ell)} \cdot M_{i,\tau_{i,\ell}^{\text{start}}-1} \mid \mathcal{G}_{t-1}\right] \\
&= \frac{1}{|A_{i,t}|}\left(\sum_{\ell \in A_{i,t-1}} \mathbb{E}\left[M_{i,t}^{(\ell)} \cdot M_{i,\tau_{i,\ell}^{\text{start}}-1} \mid \mathcal{G}_{t-1}\right] + \sum_{\ell \in A_{i,t} \setminus A_{i,t-1}} \mathbb{E}\left[M_{i,t}^{(\ell)} \cdot M_{i,t-1} \mid \mathcal{G}_{t-1}\right]\right) \\
&= \frac{1}{|A_{i,t}|}\left(\sum_{\ell \in A_{i,t-1}} M_{i,t}^{(\ell)} \cdot M_{i,\tau_{i,\ell}^{\text{start}}-1} + \sum_{\ell \in A_{i,t} \setminus A_{i,t-1}} \mathbb{E}\left[M_{i,t}^{(\ell)} \mid \mathcal{G}_{t-1}\right] \cdot M_{i,t-1}\right) \\
&\leq \frac{1}{|A_{i,t}|}\left(\sum_{\ell \in A_{i,t-1}} M_{i,t-1}^{(\ell)} \cdot M_{i,\tau_{i,\ell}^{\text{start}}-1} + \sum_{\ell \in A_{i,t} \setminus A_{i,t-1}} M_{i,t-1}\right) \\
&= \frac{1}{|A_{i,t}|}\left(|A_{i,t-1}| \cdot M_{i,t-1} + |A_{i,t} \setminus A_{i,t-1}| \cdot M_{i,t-1}\right) \\
&= M_{i,t-1}.
\end{aligned}$$

The sole inequality is because $(M_{i,t}^{(\ell)})$ is a supermartingale, and $M_{i,t}^{(\ell)} = 1$ when $t < \tau_{i,\ell}^{\text{start}}$. Thus, $\sup_{\tau \in \mathcal{T}} \mathbb{E}[E_{i,\tau}] \leq \sup_{\tau \in \mathcal{T}} \mathbb{E}[M_{i,\tau}] \leq 1$ where the final inequality is by optional stopping. Consequently, $(E_{i,t})$ are e-processes for $i \in [k]$ so $\sup_{\tau \in \mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \leq \delta$ by Fact 3, which achieves our desired result. □

Nonnegative martingales play a central role in characterizing admissible e-processes — every e-process is upper bounded by a nonnegative martingale (Corollary 24; Ramdas et al. [31]). Thus, Proposition 13 proves that if $(E_{i,t}^{(\ell)})$ are all e-processes for $i \in [k], \ell \in \mathbb{N}$, then FDR control is maintained in the multi-agent for any stopping time.

D.4 Hypotheses involving multiple arms

In the current setting, we have only considered hypotheses that are tied to a single arm i.e. hypothesis i is concerned solely with ν_i for all $i \in [k]$. We also might be concerned with hypotheses that involve multiple arms. For example, we could be interested in the hypothesis that the reward distributions are exchangeable across arms [40, 42] i.e. any permutation of the arms is the same distribution, or the hypothesis that the

means of two specific reward distributions are the same. Naturally, if each hypothesis is not restricted to being involved with only a single arm, we can consider more (or fewer) hypotheses than the number of arms.

Thus, we can denote k to be the total number of hypotheses and n to be the number of arms. Algorithm 6 specifies a meta-algorithm similar to Algorithm 1 that maintains FDR control in multi arm hypotheses. We simply maintain an e-process or p-process for each hypothesis. An important difference between hypotheses involving multiple arms setting and the standard setting is that the independence of $X_{1,t}, \dots, X_{n,t}$ is no longer sufficient to ensure all the e-variables or p-variables are dependent only through the exploration policy and stopping time. The dependence structure within the e-variables or p-variables is based not only upon the dependence of $X_{1,t}, \dots, X_{n,t}$, but also whether the hypothesis tests themselves have any dependence among each other e.g. two hypotheses might involve the same arm. Thus, for p-variables, we may require $\delta' = \delta / \log k$ even when the reward distributions are independent.

Algorithm 6: A meta-algorithm that ensures FDR control when hypotheses can involve multiple arms in the bandit setting.

Input: Exploration component $(\mathcal{A}_t)$, stopping rule τ^*, desired level of FDR control $\delta \in (0,1)$. Set $D_0 = \emptyset$.

for t *in* $1 \dots$ **do**

$\mathcal{I}_t := \mathcal{A}_t(D_{t-1}) \subseteq [n]$

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 that relate to any of the queried arms.

$$\mathcal{S}_t := \begin{cases} \text{BH}[\delta / \log k](p_{1,t}, \dots, p_{k,t}) \text{ or arbitrary p-self-consistent set} & \text{if using p-variables} \\ \text{eBH}[\delta](e_{1,t}, \dots, e_{k,t}) \text{ or arbitrary e-self-consistent set} & \text{if using e-variables} \end{cases}$$

if $\tau^* = t$ **then** stop and **return** $\mathcal{S}_t$;

end

Proposition 14. *Algorithm 6 outputs $\mathcal{S}_t$ for all $t \in \mathbb{N}$ such that $\sup_{\tau \in \mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \leq \delta$.*

E-variables in this setting have potentially larger power over p-variables than in the standard setting. This is because the number of hypotheses, k, is no longer tied to the number of arms, n. For example, $k \approx n^2/2$ if there was a hypothesis for each pair of arms in the bandit. Then, using p-variables in Algorithm 6 would require a correction of approximately $2 \log k$. In contrast, p-variables and BH require no more than a $\log k$ correction in the standard setting. Consequently, allowing for multiple arm hypotheses further highlights the benefit of e-variables over p-variables when dealing with arbitrarily dependent statistics.

E Additional simulations

In this section, we perform additional simulations to empirically verify our theoretical results. We test the performance of different choices of p-variables against e-variables in the standard bandit setting. We also provide simulations for the combinatorial bandit setting and compare p-variable methods with different assumptions against an e-variable method.

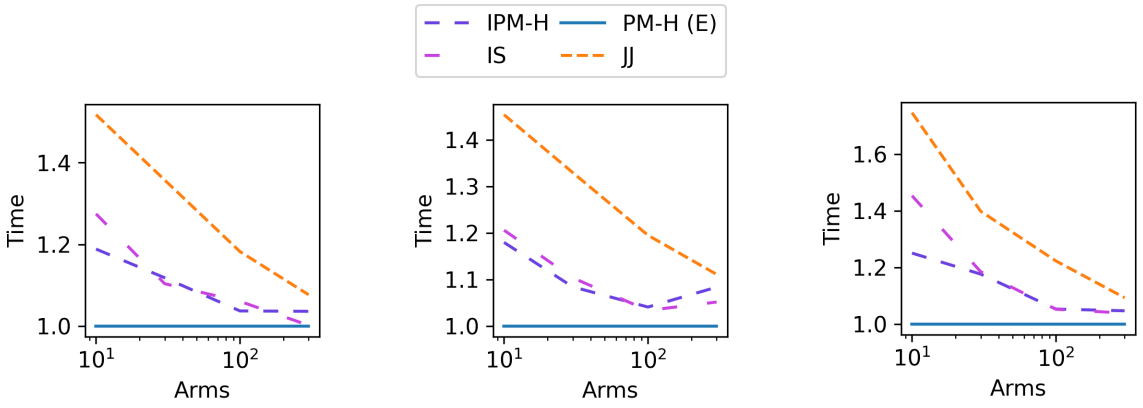
E.1 Testing against different choices of p-variables

We consider two additional choices of p-variables to compare with our e-variable method and the p-variable from JJ discussed in Section 5. One is simply $P_{i,t}^{\text{IPM-H}} := 1/E_{i,t}^{\text{PM-H}}$, which we will call Inverse PM-H (IPM-H). The other, which we call the IS p-variable, which is defined as follows by setting $\varphi = \varphi^{\text{IS}}$ in (5b).

$$P_{i,t}^{\text{IS}} := \inf\{\beta \in [0,1] : |\widehat{\mu}_{i,t} - \mu_0| > \varphi^{\text{IS}}(t, \beta)\}. \tag{15}$$

We run these methods using the UCB arm selection algorithm described in (5a) inside of Algorithm 1.

The results shown in Figure 5 demonstrate that e-variables and e-BH still perform better than any p-variable and BH method. The two new p-variables, IS and IPM-H, have about similar sample efficiency, and both outperform the JJ p-variable, but both are still slightly worse than the PM-H e-variable. Thus, e-BH and e-variables have consistently better performance than BH and p-variables.



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

Figure 5: Relative comparison of time t for each method to obtain a rejection set, $\mathcal{S}_t$, that has a $\text{TPR}(\mathcal{S}_t) \geq 1-\delta$ while maintaining $\text{FDR}(\mathcal{S}_t) \leq \delta$, where we choose $\delta = 0.05$. This plot compares different choices of p-variables against the PM-H e-variable 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 the algorithm that uses the PM-H e-variable. The JJ p-variable is the baseline p-variable specified in (5). We see that both the IPM-H and the IS p-variable have similar performance, and require fewer samples than the JJ p-variable. Overall, the PM-H e-variable performs better than any choice of p-variable.

E.2 Graph bandits with dependent $X_{1,t}, \ldots, X_{k,t}$

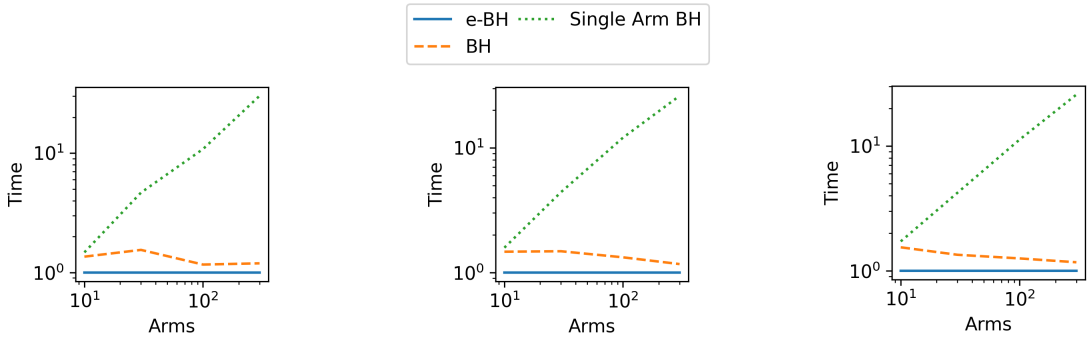
We consider a graph bandit setting where the algorithm makes no assumptions about the underlying dependence structure, and each arm consists of a node and its neighbors. We set the joint distribution over rewards at each step as the product of independent normal distributions for each arm. The marginal distribution of each arm $i \in [k]$ is a normal distribution with mean μ_i, where $\mu_i = 1/2$ if $i \in \mathcal{H}_1$ and $\mu_i = \mu_0 = 0$ if $i \in \mathcal{H}_0$. Each graph we simulate is composed of 10 cliques of $k/10$ nodes. Thus, the set of superarms available for sampling is $\mathcal{K} = \{\{i, i+10, i+20, \ldots, i+k-10\} : \text{for } i \in [10]\}$. Finally, we let $\delta = 0.05$ be level of FDR control for each algorithm.

We compare 3 different methods. For all 3 methods, the exploration strategy is to uniformly sample from the set of superarms $\mathcal{K}$. These methods differ solely in their choice of the evidence component. The first method is called the *single arm BH* method, as it only saves a single uniformly random sample from the set of samples it attains at each time step. Hence, it is equivalent to the uniformly randomly sampling BH method for the standard bandit setting. In this combinatorial bandit setting, it simply discards all but one sample at each step, and can consequently still enjoy the guarantees in Proposition 1. Our second method is to use the default BH and p-variables with no discarding of samples and the larger correction from Proposition 2. Lastly, we have the e-BH and e-variable method that also uses all samples from each pull of a superarm, since e-BH requires no correction for arbitrary dependence.

Figure 6 shows the results of using methods that guarantee FDR control at level δ on graph bandits with arbitrary dependence between arms. Single arm BH pays a tremendous cost in time by throwing away many samples at each step, and the slightly smaller correction it needs to make does not make up for this deficit. Between the two methods that make full use of the samples obtained from superarm, we see that e-BH does better. Thus, e-variables and e-BH exhibit empirical performance on par or better than p-variables and BH in both the standard and combinatorial bandit settings.

F Betting interpretation of e-variables for bandits

We will describe our methodology for constructing e-variables using the perspective of betting in this section. Shafer [33] uses betting to formulate a paradigm for understanding the quantity represented by an e-value, and Shafer and Vovk [34] extend these ideas to form a mathematically rigorous foundation for probability based on game theory. Separately, betting ideas have also been used in parameter free techniques for online



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

Figure 6: Relative comparison of time t for each method to obtain a rejection set, $\mathcal{S}_t$, that has a $\text{TPR}(\mathcal{S}_t) \geq 1-\delta$ while maintaining $\text{FDR}(\mathcal{S}_t) \leq \delta$, where we choose $\delta = 0.05$. This plot compares two different p-variable methods (BH and single arm BH) against an e-variable method (e-BH) 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 the algorithm that uses e-BH. We see that e-BH outperforms the two other BH algorithms in the graph bandit setting. Notably, single arm BH is linearly increasing in time relative to the other two methods that make full use of the samples obtained from a superarm. Single arm BH discards too many samples at each step, and the smaller correction it makes does not make up the deficit in number of samples.

learning [28, 21, 29]. In this section, we will use a betting approach to produce a data adaptive e-process.

Recall that if E is an e-variable, then $\mathbb{E}[E] \leq 1$ when the null hypothesis is true. On the other hand, if the null hypothesis is false, we would like E to be large, since that increases the likelihood that the null hypothesis is rejected. Thus, constructing E such that is satisfies the e-variable constraint under the null and is large under the alternative is the same as constructing a valid hypothesis test that has as much power as possible. Consequently, we can consider a betting game where we pay a dollar to play, and E is the payout. If the null hypothesis is true, then we are unable to make any money in expectation, since the expectation is of E is at most 1. However, if the null hypothesis is false, then we would expect to be able to make money on this game. If we did not make money under the alternative, then any test that used this e-variable would have no power, since the behavior of E would not change between the null hypothesis being true and being false. In other words, this would be no better than picking $E = 1$ deterministically: a valid e-variable, but ineffectual for testing.

We define the **predictably-mixed Hoeffding (PM-H)** e-process [45], which we used in our simulations in Section 5, as follows:

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

$(\lambda_{i,t})$ is any sequence of nonnegative real numbers that is predictable w.r.t. $(\mathcal{F}_t)$. We will use an argument based on betting to derive a $(\lambda_{i,t})$ sequence, and show that this e-process can "make money" and hence provide TPR guarantees in the sub-Gaussian case. We first observe the following property of this process.

Proposition 15. *$E_{i,t}^{\text{PM-H}}(\mu_0)$ is a nonnegative supermartingale, and thus an e-process, if $i \in \mathcal{H}_0$ and ν_i is 1-sub-Gaussian.*

Proof. We drop μ_0 from $(E_{i,t}^{\text{PM-H}}(\mu_0))$ and denote it as $(E_{i,t}^{\text{PM-H}})$.

We proceed by showing $(E_{i,t}^{\text{PM-H}})$ is a nonnegative supermartingale w.r.t. to the canonical filtration $(\mathcal{F}_t)$.

Consider $E_{i,t}^{\text{PM-H}}$ when $i \in \mathcal{H}_0$. If $I_t \neq i$ then $E_{i,t}^{\text{PM-H}} = E_{i,t-1}^{\text{PM-H}}$, which satisfies the supermartingale property in (6).

Otherwise,

$$\begin{aligned}\mathbb{E}[E_i^{\text{PM-H}} i, t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\lambda_t(X_{i,t}-\mu_0) - \frac{\lambda_t^2}{2}\right) E_i^{\text{PM-H}} i, t-1 \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\exp\left(\lambda_t(X_{i,t}-\mu_0) - \frac{\lambda_t^2}{2}\right) \mid \mathcal{F}_{t-1}\right] E_{i,t-1}^{\text{PM-H}} \\ &\leq E_{i,t-1}^{\text{PM-H}},\end{aligned}$$

where the final equality is because $X_{i,t}$ are independent across $t \in \mathbb{N}$ and 1-sub-Gaussian, and $\mu_i \leq \mu_0$.

Since $(E_{i,t}^{\text{PM-H}})$ is a nonnegative supermartingale, it is an e-process by optional stopping. Thus, we have achieved our desired result. □

Note that Proposition 15 justifies that our choice of e-process for the simulations in Section 5 was indeed a valid e-process. Now that we have shown $(E_{i,t}^{\text{PM-H}})$ is an e-process, we will consider how to choose a powerful $(\lambda_{i,t})$. Consider a model where we view the e-value, $e_{i,t}$, for each arm $i \in [k]$ as the money made by each arm, or a "**betting score**". For each arm $i \in [k]$, imagine we are allocated initial wealth equal to 1. At each time step, the algorithm chooses an arm $i \in [k]$, and a "bet", $\lambda_{i,t}$. The wealth of the arm at the next round changes by a factor based on the reward $X_{i,t}$ (assuming i is the arm chosen at round $t+1$):

$$e_{i,t+1} = e_{i,t} \cdot \underbrace{\exp(\lambda_{i,t}(X_{i,t}-\mu_0) - \lambda_{i,t}^2/2)}_{\text{change in wealth}}.$$

Note that this a "fair game" or the reward multiplier is less than 1 in expectation if $\mathbb{E}[X_{i,t}] \leq \mu_0$.

The betting score, $E_{i,t}^{\text{PM-H}}$, may be interpreted as the money earned by arm i at time t. When the null hypothesis is true, i.e. $\mu_i \leq \mu_0$, we know that $\sup_{\tau \in \mathcal{T}} \mathbb{E}[E_{i,\tau}^{\text{PM-H}}] \leq 1$ by Proposition 15. Thus, regardless of our stopping strategy, we make no money in expectation. However, if we knew that $E_{i,t}^{\text{PM-H}}$ was actually a favorable bet, and $\mathbb{E}[X_{i,t}] = \mu_i > \mu_0$, we would want to come up with a sequence $(\lambda_{i,t})$ for each arm $i \in [k]$ that maximizes our wealth at each arm. Consequently, we can reframe our goal for choosing $(\lambda_{i,t})$ as maximizing capital in a betting game. In the next section, we will discuss some strategies for accomplishing such an objective.

F.1 Optimal betting strategies

One way of maximizing capital is to optimize for the Kelly criterion [24], which aims to maximize the logarithm of the capital on each step and is equivalent to maximizing rate of growth of capital. In our scenario, the Kelly criterion manifests in the following form:

$$\mathbb{E}\left[\log E_{i,t}^{\text{PM-H}}(\mu_0)\right] = \sum_{j=1}^{T_i(t)} \mathbb{E}\left[\lambda_{i,t_i(j)}(X_{i,t_i(j)}-\mu_0) - \lambda_{i,t_i(j)}^2/2\right].$$

Optimal choice of (λ_t) for log wealth. To maximize the above sum, we can simply decompose it with respect to each j, and since the $\lambda_{i,t_i(j)}$ are decoupled, we can identify an optimal $\lambda_{i,t_i(j)}^*$ for each j:

$$\begin{aligned}\lambda_{i,t_i(j)}^* &:= \operatorname{argmax}_{\lambda \in \mathbb{R}^+} \ \lambda \mathbb{E}[X_{i,t_i(j)} - \mu_0] - \lambda^2/2 = \mu_i, \\ \lambda_{i,t_i(j)}^* \mathbb{E}[X_{i,t_i(j)} - \mu_0] - {\lambda_{i,t_i(j)}^*}^2/2 &= \max_{\lambda \in \mathbb{R}^+} \ \lambda \mathbb{E}[X_{i,t_i(j)} - \mu_0] - \lambda^2/2 = \Delta_i^2/2.\end{aligned}$$

We can see that if the μ_i is known, the above quantity is maximized by setting $\lambda_{i,t_i(j)} = \mu_i$ for all $j \in [T_i(t)]$. This observation confirms our intuition that the Kelly criterion is a sensible quantity to optimize for when trying to maximize the e-values of hypothesis in $\mathcal{H}_1$. On the other hand, if $i \in \mathcal{H}_0$, $\mu_i \leq \mu_0 = 0$, the log wealth incurred at each time step is nonpositive. Thus, in expectation, the log wealth process $\log E_{i,t}^{\text{PM-H}}(\mu_0)$ will only increase in capital when the hypothesis associated with the arm is truly non-null. In betting language, we are presenting a one-sided bet that allows for our bets $(\lambda_{i,t})$ to make money in

expectation iff the null hypothesis is false. Hence, our strategy is profitable only when the true mean of the arm is greater than μ_0.

In practice, we do not know μ_i, since testing μ_i is the entire premise of the problem. Instead, we can use the sample mean, $\widehat{\mu}_{i,t}$, in place of μ_i and show that it gives us convergence at a rate of approximately $1/T_i(t)$ to the optimal capital gain rate.

Proposition 16. *Let ν_i be 1-sub-Gaussian for $i \in [k]$. If $\lambda_{i,t} = \hat{\mu}_{i,t-1}$, then*

$$\mathbb{E}[\widehat{\mu}_{i,t_i(j)-1}(X_{i,t_i(j)} - \mu_0) - \widehat{\mu}^2_{i,t_i(j)-1}/2] = \Delta_i^2/2 - 1/(T_i(t) - 1).$$

Proposition 16 follows from the variance of $\widehat{\mu}_{i,t}$ being $1/T_i(t)$. Now, we can derive the following corollary.

Corollary 1. *The total log wealth at time t, $\log E^{\text{PM-H}}_{i,t}$, has an expectation satisfying the following property, where $\lambda_{i,t} = \widehat{\mu}_{i,t-1}$:*

$$\mathbb{E}[\log E^{\text{PM-H}}_{i,t}] = T_i(t)\Delta_i^2/2 - \sum_{j=1}^{T_i(t)} 1/j \approx T_i(t)\Delta_i^2/2 - \log(T_i(t)),$$

where $\sum_{j=1}^{T_i(t)} 1/j$ is approximately $\log(T_i(t))$.

Thus, in log wealth, using $\widehat{\mu}_{i,t}$ incurs a penalty of $\log T_i(t)$, which is relatively small compared to the positive term — especially when t is large.

F.2 Sample complexity for standard sub-Gaussian bandits

We prove a sample complexity result for the $E^{\text{PM-H}}_{i,t}$ as well.

Theorem 2. *Let $(\mathcal{A}_t)$ be such that $\mathcal{A}_t$ outputs $\mathcal{I}_t = \{I_t\}$ for all $t \in \mathbb{N}$, where I_t is defined in (5a), $\lambda_{i,t} = (\widehat{\mu}_{i,t-1}/2)_+$, and $E_{i,t} = E^{\text{PM-H}}_{i,t}$. Then, Algorithm 1 will always guarantee $\sup_{\tau\in\mathcal{T}} \text{FDR}(\mathcal{S}_\tau) \le \delta$. With at least $1-\delta$ probability, there will exist*

$$T \lesssim \sum_{i\in\mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \sum_{i\in\mathcal{H}_1} \Delta_i^{-2}(\log(\Delta_i^{-2})\log(1/\delta) + \log k)$$
$$\wedge\, |\mathcal{H}_0|\Delta^{-2}\log(\log(\Delta^{-2})/\delta) + |\mathcal{H}_1|\Delta^{-2}\log(\Delta^{-2})\log(1/\delta)$$

such that $\text{TPR}(\mathcal{S}_t) \ge 1-\delta$ for all $t \ge T$.

Theorem 2 shows the limitation of using an estimate of the mean, $\widehat{\mu}_{i,t}$, in place of the true mean. The intuition of the proof of Theorem 2 is that at each step, $E^{\text{PM-H}}_{i,t}$ must account for an $1/t$ deviation, since the variance of $\widehat{\mu}_{i,t}$ is $1/t$. The sum of these deviations is approximately $\log t$. Thus, the sample complexity bound has a $\log \Delta^{-2}$ instead of only a $\log\log \Delta^{-2}$ term. This limitation seems to be an inherent flaw in choice of $(\lambda_{i,t})$ based on estimation, since the estimation error must be accounted for along with the typical deviation from providing a concentration inequality that is uniform over time steps t.

To prepare for our proof of Theorem 2, we require some self-contained lemmata. Define the following auxiliary random variables for all $i \in \mathcal{H}_1$:

$$\rho_i' := \min_{t\in\mathbb{N}} \frac{E_{i,t}}{\exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}\Delta_i - \lambda^2_{i,t_i(j)}\right)}. \tag{16}$$

Lemma 7. *For all $i \in \mathcal{H}_1$, $\mathbb{P}(\rho_i' \le s) \le s$ for $s \in (0,1)$ i.e. ρ_i' is superuniformly distributed.*

Proof. We observe that the reciprocal of ρ_i' is the following:

$$1/\rho_i' = \max_{t\in\mathbb{N}} \exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}(\mu_i - X_{i,t_i(j)}) - \lambda_{i,t_i(j)}^2/2\right).$$

Let

$$M_t = \exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}(\mu_i - X_{i,t_i(j)}) - \lambda_{i,t_i(j)}^2/2\right).$$

We will show (M_t) is a nonnegative supermartingale w.r.t. $(\mathcal{F}_t)$. Assume arm i is sampled at time t — otherwise the supermartingale property is trivially satisfied.

$$\begin{aligned}
\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}(\mu_i - X_{i,t_i(j)}) - \lambda_{i,t_i(j)}^2/2\right) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\exp\left(\lambda_{i,t}(\mu_i - X_{i,t}) - \lambda_{i,t}^2/2\right) \mid \mathcal{F}_{t-1}\right] \exp\left(\sum_{j=1}^{T_i(t-1)} \lambda_{i,t_i(j)}(\mu_i - X_{i,t_i(j)}) - \lambda_{i,t_i(j)}^2/2\right) \\
&\le \exp\left(\sum_{j=1}^{T_i(t-1)} \lambda_{i,t_i(j)}(\mu_i - X_{i,t_i(j)}) - \lambda_{i,t_i(j)}^2/2\right) \\
&= M_{t-1}.
\end{aligned}$$

The sole inequality arises from $X_{i,t}$ being independent across $t \in \mathbb{N}$ and 1-sub-Gaussian, and having mean μ_i.

Thus, ρ_i' is superuniformly distributed by Ville's inequality. □

Rewriting the definition of ρ_i', we get:

$$E_{i,t} \ge \exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}\Delta_i - \lambda_{i,t_i(j)}^2\right)\rho_i' \tag{17}$$

for all $t \in \mathbb{N}$. Now show a result for the rate of growth of $E_{i,t}$ by showing a result concerning the lower bound in (17).

Lemma 8. *For all $t \in \mathbb{N}$,*

$$\exp\left(\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}\Delta_i - \lambda_{i,t_i(j)}^2\right) \gtrsim T_i(t)\Delta_i^2 - \log(1/\rho_i)\log(T_i(t)).$$

Proof. Recall that $\lambda_t = (\widehat{\mu}_{i,t-1}/2)_+$. Then, we derive the following asymptotic lower bound:

$$\begin{aligned}
\sum_{j=1}^{T_i(t)} \lambda_{i,t_i(j)}\Delta_i - \lambda_{i,t_i(j)}^2 &= \frac{1}{4}\sum_{j=1}^{T_i(t)} 2\widehat{\mu}_{i,t-1}\Delta_i - \widehat{\mu}_{i,j-1}^2 \\
&\geq \frac{1}{4}\sum_{j=1}^{T_i(t)} \Delta_i^2 - (\Delta_i - \widehat{\mu}_{i,j-1})^2 \\
&\geq \frac{1}{4}\sum_{j=1}^{T_i(t)} \Delta_i^2 - \varphi(T_i(j-1),\rho_i)^2 && \text{def. of } \rho_i \\
&\geq \frac{1}{4}\sum_{j=1}^{T_i(t)} \Delta_i^2 - \frac{4\log(\log_2(2j)/\rho_i)}{j} && \text{upper bound from Fact 4} \\
&\gtrsim \Delta_i^2 T_i(t) - \log\left(\frac{1}{\rho_i}\right)\log(T_i(t)),
\end{aligned}$$

where the last line is because $\sum_{j=1}^{T_i(T)} 1/j \approx \log T_i(t)$. Thus, we have arrived our desired result. □

We now have the ingredients to present a proof of Theorem 2.

Proof of Theorem 2. Combining the lower bound in (17) with Lemma 8, we get the following asymptotic lower bound:

$$E_{i,t} \gtrsim \exp(T_i(t)\Delta_i^2 - \log(1/\rho_i') - \log(1/\rho_i)\log(T_i(t)).$$

Inverting the expression above, we get that the following lower bound sample complexity of a single arm,

$$T_i(t) \gtrsim \Delta_i^{-2}\log(\Delta_i^{-2})\log(1/\rho_i) + \Delta_i^{-2}\log(1/\rho_i') + \Delta_i^{-2}\log(\varepsilon),$$

implies $E_{i,t} \geq \varepsilon$ for $\varepsilon > 0$.

We can now derive a bound for $\sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\}$.

$$\sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\} \leq \max_{\pi} \sum_{i\in\mathcal{H}_1} \Delta_i^{-2}\log(\Delta_i^{-2})\log(1/\rho_i) + \Delta_i^{-2}\log(1/\rho_i') + \Delta_i^{-2}\log(k/\pi(i))$$

where π is a mapping from $[|\mathcal{H}_1|]$ to $\mathcal{H}_1$.

We get the following total bound:

$$\begin{aligned}
\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} &= \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} + \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\} \\
&= \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_0, \mathcal{R} \not\subseteq \mathcal{S}_t\} + \sum_{t=1}^{\infty} \mathbf{I}\{I_t \in \mathcal{H}_1, \mathcal{R} \not\subseteq \mathcal{S}_t\} \\
&\lesssim \sum_{i\in\mathcal{H}_0} \Delta_i^{-2}\log(\log(\Delta_i^{-2})/\delta\rho_i) \\
&\quad + \max_{\pi} \sum_{i\in\mathcal{H}_1} \Delta_i^{-2}\log(\Delta_i^{-2})\log(1/\rho_i) + \Delta_i^{-2}\log(1/\rho_i') + \Delta_i^{-2}\log(k/\pi(i)),
\end{aligned}$$

where the asymptotic inequality is by Lemma 5.

We know that we can apply Lemma 4 at level $\beta = \delta/2$ to ρ_i for $i \in [k]$ and ρ_i' for $i \in \mathcal{H}_1$. Thus, the following happens with at least $1 - \delta$ probability:

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \max_{\pi} \sum_{i \in \mathcal{H}_1} \Delta_i^{-2} \log(\Delta_i^{-2}) \log(1/\delta) + \Delta_i^{-2} \log(k/\pi(i)).$$

Similar to the Theorem 1, we can show two different bounds. The first is the following:

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim |\mathcal{H}_0| \Delta^{-2} \log(\log(\Delta^{-2})/\delta) + |\mathcal{H}_1| \Delta^{-2} \log(\Delta^{-2}) \log(1/\delta),$$

because $\sum_{i=1}^{|\mathcal{H}_1|} \log(k/i) \leq k$. The second follows from dropping $\pi(i)$:

$$\sum_{t=1}^{\infty} \mathbf{I}\{\mathcal{R} \not\subseteq \mathcal{S}_t\} \lesssim \sum_{i \in \mathcal{H}_0} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta) + \sum_{i \in \mathcal{H}_1} \Delta_i^{-2} \log(\Delta_i^{-2}) \log(1/\delta) + \Delta_i^{-2} \log k.$$

Thus, we have shown both of our desired bounds. □

G Testing the average conditional mean

For simplicity, we will discuss results and proofs in this section under the single arm bandit case, so we will drop the arm index i when labeling terms. Our conclusions, however, do generalize to the general multi-arm bandit case.

In Section 4 and JJ, the null hypothesis for each arm we are concerned with is

$$\text{“}\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq \mu_0 \text{ for all } t \in \mathbb{N} \text{ almost surely.”} \tag{H1}$$

In the aforementioned settings, there is an additional assumption that $X_{i,t}$ are i.i.d. across $t \in \mathbb{N}$. Thus, $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ simply becomes $\mathbb{E}[X_t]$. We can also test a more general hypothesis of whether the means of X_t are less than or equal to μ_0 on average.

To formally define a notion of “average mean”, let us consider the case where there is a single arm i.e. we have a sequence of rewards $X_1, X_2, \ldots$, where the *average conditional mean* is defined as

$$\overline{\mu}_t \equiv \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}[X_j \mid \mathcal{F}_{j-1}].$$

Consequently, we can define a null hypothesis w.r.t. $\overline{\mu}_t$:

$$\text{“}\overline{\mu}_t \leq \mu_0 \text{ for all } t \in \mathbb{N} \text{ almost surely.”} \tag{H2}$$

In the specific case where X_t are i.i.d. across $t \in \mathbb{N}$, each with mean μ, then $\mathbb{E}[X_t \mid \mathcal{F}_{j-1}] = \mathbb{E}[X_t] = \mu$ for all $t \in \mathbb{N}$, and $\overline{\mu}_t = \mu$. Consequently, there would be no difference between testing the average conditional mean and testing the marginal mean, μ, because they are the same value. However, when the distribution of X_t are not necessarily i.i.d. across $t \in \mathbb{N}$, we will emphasize that not all valid tests for (H1) are also valid for (H2). Generally, (H1) is a “stronger” hypothesis than (H2) in the sense that any distribution over X_t for $t \in \mathbb{N}$ that satisfies (H1) also satisfies (H2).

The difference between (H1) and (H2) is reflected in the fact that e-processes are supermartingales in (H1), but only upper bounded by a martingale in (H2).

Proposition 17. *Assume that conditional distribution of $X_t \mid \mathcal{F}_{t-1}$ is always 1-sub-Gaussian for all $t \in \mathbb{N}$. Consider a process of the form,*

$$E_t := \sum_{\ell=1}^{m} w_\ell \exp\left(\sum_{j=1}^{t} \lambda_\ell (X_j - \mu_0) - \frac{\lambda_\ell^2}{2}\right)$$

where $m \in \mathbb{N} \cup \{\infty\}$, $\sum_{\ell=1}^{m} w_\ell \leq 1$. Under both (H1) and (H2), (M_t) is a e-process. Specifically, (M_t) is

(i) a nonnegative supermartingale under (H1).

(ii) upper bounded by a nonnegative supermartingale under (H2).

Proof. (i) follows from Proposition 7.

Without loss of generality, we will consider the case where $m = 1$ and $w_1 = 1$, since a convex combination of supermartingales is a supermartingale. Thus,

$$E_t = \exp\left(\sum_{j=1}^{t} \lambda (X_j - \mu_0) - \frac{\lambda^2}{2}\right).$$

To prove (ii), we first notice we can define a process M_t' that upper bounds E_t:

$$\begin{aligned}
M_t' &= \exp\left(\sum_{j=1}^{t} \lambda (X_j - \mathbb{E}[X_j \mid \mathcal{F}_{j-1}]) - \frac{\lambda^2}{2}\right) \\
&= \exp\left(\lambda \left(\sum_{j=1}^{t} X_j - \sum_{j=1}^{t} \mathbb{E}[X_j \mid \mathcal{F}_{j-1}]\right) - \frac{t\lambda^2}{2}\right) \\
&= \exp\left(\lambda \left(\sum_{j=1}^{t} X_j - t\overline{\mu}_t\right) - \frac{t\lambda^2}{2}\right) \\
&\geq \exp\left(\lambda \left(\sum_{j=1}^{t} X_j - t\mu_0\right) - \frac{t\lambda^2}{2}\right) \\
&= E_t.
\end{aligned}$$

Now, we will show that (M_t') is a supermartingale.

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{j=1}^{t} \lambda (X_j - \mathbb{E}[X_j \mid \mathcal{F}_{j-1}]) - \frac{\lambda^2}{2}\right) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\exp\left(\lambda (X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]) - \frac{\lambda^2}{2}\right) \mid \mathcal{F}_{t-1}\right] M_{t-1}' \\
&\leq M_{t-1}',
\end{aligned}$$

where the last inequality is because the conditional distribution of $X_t \mid \mathcal{F}_{t-1}$ is 1-sub-Gaussian. Thus, we have shown both parts of our desired result. □

The E_t specified in Proposition 17 is an e-process, but not necessarily a nonnegative supermartingale, for any distribution under (H2) where there exists a $t \in \mathbb{N}$ such that $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] > \mu_0$. Thus, the distinction highlighted in Proposition 17 is not vacuous.

Further, we will also note the following negative result that there exists processes that are e-processes under (H1) but are not under (H2).

Proposition 18. *Assume that conditional distribution of $X_t \mid \mathcal{F}_{t-1}$ is always 1-sub-Gaussian for all $t \in \mathbb{N}$. $(E_t^{\text{PM-H}})$ is an e-process under all distributions satisfying (H1), but there exist (λ_t) such that $(E_t^{\text{PM-H}})$ is not an e-process under all distributions that satisfy (H2).*

Proof. $(E_t^{\text{PM-H}})$ is an e-process under (H1) by a similar argument to the proof of Proposition 15, since the conditional distribution of $X_t \mid \mathcal{F}_{t-1}$ is 1-sub-Gaussian. However, we can provide a simple counterexample choice of (λ_t) and distribution that satisfies (H2) which cannot have expectation greater than 1 at a time $t \in \mathbb{N}$. Let $\mu_0 = 0$, $X_t = -1$ if t is odd, and $X_t = 1$ if t is even. Consider a (λ_t) where $\lambda_t = 0$ when t is odd and $\lambda_t = 1$ when t is even. Then,

$$\mathbb{E}[E_t^{\text{PM-H}}] = \exp(\lceil t/2 \rceil).$$

Consequently, $(E_t^{\text{PM-H}})$ with this choice of (λ_t) is not an e-process under (H2), and we have proved our desired result. □

Thus, using adaptive strategies for selecting (λ_t) like in Waudby-Smith and Ramdas [45] for testing (H2) is not necessarily straightforward, while mixture strategies in the form specified in Proposition 17 are valid e-processes for testing both (H1) and (H2).