Optimal Stopping Rules for Best Arm Identification in Stochastic Bandits under Uniform Sampling

Vedang Gupta†, Yash Gadhia†, Shivaram Kalyanakrishnan, Nikhil Karamchandani††
Indian Institute of Technology Bombay
200100166@iitb.ac.in, gadhiayash@gmail.com, shivaram@cse.iitb.ac.in, nikhilk@ee.iitb.ac.in

***Abstract*—We consider the problem of best arm identification in stochastic multi-armed bandits, in the setting that each arm is sampled once in each round. This *uniform sampling* regime is a conceptually simple setting that is relevant to many practical applications. The aim is to stop and correctly identify the best arm with probability at least $1-\delta$, while keeping the number of rounds low. We derive a lower bound on the sample complexity for this setting. Thereafter, we propose two natural stopping rules for Bernoulli bandits: one based on PPR martingale confidence sequences, and the other based on the GLR statistic. Both rules are shown to match the lower bound as $\delta \to 0$. Our analysis and experiments suggest that the relative performance of the two stopping rules depends on a property of the bandit instance.**

I. Introduction

We consider the problem of Best Arm Identification (BAI) in stochastic multi-armed bandits. In contrast with the classical task of regret-minimisation [1,2], BAI is a problem of "pure exploration". The aim is to identify the most rewarding (or *best*) arm from a set by sampling the arms, without suffering any explicit penalty for pulling inferior arms. Pure exploration is relevant when experiments are conducted off-line. BAI in particular finds a variety of applications, including reinforcement learning [3], clinical trials [4], recommendation systems [5], crowdsourcing [6,7], brain-computer interfaces [8], and Monte Carlo tree search [9].

BAI has been studied in two main settings. In the **fixed budget** setting, the algorithm is constrained to limit its experimentation to a given budget of T pulls [10]. The goal is to minimise either the probability of mis-identifying the best arm, or a related quantity called the "simple regret". Literature from the last decade has reduced the gap between upper and lower bounds for BAI in the fixed budget setting [10]–[12], although the gap remains open [13]. The second setting of BAI is that of **fixed confidence**, wherein the input to the algorithm is a mistake probability δ, and the aim is to minimise the number of pulls to guarantee that the probability of error does not exceed δ [3]. Our investigation is in the fixed confidence setting, which has received extensive attention in the literature.

Among the earliest algorithms for BAI in the fixed confidence setting are ones that sequentially eliminate arms, until only the winner remains [3,11]. A common aspect of several algorithms in this setting is for sampling and stopping both to be guided by lower and upper confidence bounds on the unknown means of the arms [14]–[16]. Such algorithms enjoy sample complexity upper bounds that depend on the problem hardness, and are typically within a constant factor of the dominant δ-dependent term in the lower bound.

Garivier and Kaufmann [17] ushered in a significant shift in the analysis of algorithms in the fixed confidence setting. These authors proposed the notion of *asymptotic optimality*; the ratio of the sample complexity of their "track and stop" algorithm to the applicable lower bound approaches 1 as $\delta \to 0$. At the core of their algorithm is a calculation of the fraction of pulls each arm must receive; unfortunately this is an expensive numerical computation to be performed after each pull. A more computationally feasible alternative is presented in the form of several Bayesian algorithms, which choose probabilistically between an estimated best arm and a challenger at each round. Among such algorithms are "Top-Two Thompson Sampling" [18], "Top-Two Transportation Cost" [19], and "BayesElim" [20].

A. Our Contribution

The algorithms described thus far are all "fully sequential", in that the decision of which arm to pull next is recomputed after every single pull (or a *constant* number of pulls). In practice, the experimenter may not have the ability to continuously monitor samples and carefully readjust the allocation of samples. Moreover, in some applications, it is actually possible to simultaneously sample multiple or even all the arms of the bandit instance without paying an additional price. Examples include computer simulations on a parallel cluster [21], and surveys conducted across multiple geographical locations [22].

We study BAI in the fixed confidence setting, in the regime of *uniform sampling*. Simply put, the learning algorithm receives a fresh sample for each arm in every round; the only decision to make is when to stop (at termination, it is arguably optimal to return the empirically-best arm) as opposed to the bandit setting where one also needs an appropriate sampling rule. We reuse the template proposed by Kaufmann et al. [23] to work out a lower bound on the round complexity for this problem (Section III). We then propose two separate stopping rules for Bernoulli bandits. The first rule, denoted **PPR-JD**, is based on PPR martingale confidence sequences [24], which were recently also applied to the closely-related problem of PAC mode estimation [25]. The second rule, denoted **U-CNF**, is based on the Chernoff rule [17]. After presenting these stopping rules in Section IV, we show that indeed both are

†V. Gupta and Y. Gadhia were partially supported by C-MInDS IIT Bombay. ††N. Karamchandani's work was supported by a SERB grant on 'Online Learning with Constraints' and a SERB MATRICS grant.

asymptotically optimal in their sample complexity (while also being computationally efficient). We validate our analytical findings through experiments in Section V.

II. Problem Statement

We consider a stochastic bandit with $K \geq 2$ arms, and denote the set of arms by $[K] \coloneqq \{1, 2, \ldots, K\}$. Each arm $a \in [K]$ has an associated distribution Π_a over scalar rewards, which is *a priori* unknown to the learner. When arm a is pulled, it earns a random reward $r \sim \Pi_a$. Rewards from the same arm a are i.i.d. samples from Π_a. We denote by μ_a the mean of Π_a: in other words, if $r \sim \Pi_a$, then $\mathbb{E}[r] = \mu_a$. In order that the problem of best arm-identification be well-defined, we assume that any given bandit instance has a unique best arm $a^* \in [K]$. Without loss of generality, we can index the arms in non-ascending order of their mean rewards. Thus,

$$\mu_1 > \mu_2 \geq \mu_3 \geq \mu_4 \geq \cdots \geq \mu_K. \tag{1}$$

In best arm identification (BAI), the goal of the learner is to correctly identify the arm with the highest mean using as few samples as possible. The interaction between the learning algorithm and the bandit instance proceeds in rounds. In an unrestricted setup, the algorithm can specify an arbitrary, single arm to be pulled in each round [3,23]. However, in this paper, we consider the regime of *uniform sampling*, where in each round, every arm is pulled exactly once. Devoid of any need for decision making when it comes to sampling, the algorithm may simply be viewed as a *stopping rule*, which at the end of each round decides whether (1) to stop and declare an estimate $\hat{a}$ for the best arm, or (2) to perform another round of pulls. The input to the algorithm at the beginning of each round is the history of outcomes registered thus far. The indexing of the means in (1) is not available to the algorithm; it is only used for our analysis.

For $\delta \in (0, 1]$, algorithm $\mathcal{A}$ is said to be δ-PAC if on every bandit instance $\mathcal{I}$, $\mathbb{P}_{\mathcal{I}}(\hat{a} \neq a^*) \leq \delta$. In other words, δ upper-bounds the mistake probability of $\mathcal{A}$ on every bandit instance $\mathcal{I}$. For a given instance $\mathcal{I}$, let random variable $N_{\delta,\mathcal{I}}$ denote the number of rounds taken by $\mathcal{A}$ to terminate when run on $\mathcal{I}$. Informally, our goal is to construct δ-PAC strategies for which $N_{\delta,\mathcal{I}}$ is "small". Formally, we seek δ-PAC algorithms that minimise $\mathbb{E}[N_{\delta,\mathcal{I}}]$, which is the round complexity and K times the sample complexity. Before proceeding, we introduce some notation that will be used in the upcoming sections.

A. Notation

A bandit instance $\mathcal{I}$ fixes the probability distribution Π_a of each arm $a \in [K]$. We denote by $\mathrm{KL}(P \parallel Q)$, the Kullback-Leibler divergence between the two probability distributions P and Q. Denote by Ω the set of all bandit instances that have each arm's reward distribution drawn from the set of probability measures $\mathit{\Pi}$ satisfying, $0 < \mathrm{KL}(P \parallel Q) < \infty$ for $P, Q \in \mathit{\Pi}$, $P \neq Q$. Such a set of bandit models Ω is called *identifiable* [23]. Denote by $\mathcal{B}$ the set of all bandit instances where each arm's reward distribution is Bernoulli. Since bandit instances in $\mathcal{B}$ can be fully parameterised by their means, we denote such a bandit instance by $\bar{\mu} \coloneqq (\mu_1, \mu_2 \ldots, \mu_K)$, which defines the instance $\Pi_a = \text{Bernoulli}(\mu_a)$, for all $a \in [K]$. We also specifically define the term:

$$d(x, y) \coloneqq x \log\left(\frac{x}{y}\right) + (1 - x) \log\left(\frac{1 - x}{1 - y}\right)$$

along with the additional convention that $d(0, 0) = d(1, 1) = 0$, which ensures $d(x, y) = 0$ whenever $x = y$. Note that $d(\mu_a, \mu_b) = \mathrm{KL}(\text{Bernoulli}(\mu_a) \parallel \text{Bernoulli}(\mu_b))$ is the relative entropy between two Bernoulli distributions. All logarithms in this paper are natural logarithms.

We shall use (i) t for counting the overall *time*—that is, the number of pulls up to that point, (ii) n for counting the number of rounds, (iii) τ for the stopping time, and (iv) N for the stopping round. In the case of uniform sampling, observe that $\tau = NK$.

In our algorithms, we shall use $\hat{\mu}_a$ to represent the empirical mean of arm a up to the *current* time τ or round n (implicit from context). We also maintain an ordering of the empirical means of the arms satisfying $\hat{\mu}_{\alpha_1} \geq \hat{\mu}_{\alpha_2} \geq \cdots \geq \hat{\mu}_{\alpha_K}$, where α_i denotes the arm with the i^{th} highest empirical mean.

III. Lower Bounds on the Round Complexity

In this section, we present lower bounds on the round complexity of δ-PAC algorithms under uniform sampling. We follow the same sequence of steps as Kaufmann et al. [23] for the unrestricted setting, while making suitable modifications to account for uniform sampling.

Theorem 1 (General Lower Bound): Let $\mathcal{I} \in \Omega$ be any identifiable bandit instance with a unique best arm. Given a mistake probability $\delta \in (0, 1)$, any δ-PAC uniform sampling algorithm on $\mathcal{I}$ with stopping round $N_{\delta,\mathcal{I}}$ satisfies

$$\mathbb{E}[N_{\delta,\mathcal{I}}] \geq \frac{\log(1/2.4\delta)}{\inf\limits_{\mathcal{I}' \in \Omega'(\mathcal{I})} \sum\limits_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a)}, \tag{2}$$

where $\Omega'(\mathcal{I}) = \{\mathcal{I}' \mid a^*(\mathcal{I}') \neq a^*(\mathcal{I}), \mathcal{I}' \in \Omega\}$ is the set of all identifiable bandit instances having a best arm that is different from that of $\mathcal{I}$.

The above result follows immediately from [23, Lemma 1] and by noting that under uniform sampling, the number of samples is the same for all arms under any stopping rule. We include the details in Appendix A for completeness. For Bernoulli bandits, we obtain a closed form expression for the above lower bound.

Corollary 2 (Lower Bound for Bernoulli Bandits): Let $\bar{\mu}$ be an identifiable Bernoulli bandit instance with a unique best arm. Given a mistake probability $\delta \in (0, 1)$, any δ-PAC uniform sampling algorithm on $\bar{\mu}$ with stopping round $N_{\delta,\bar{\mu}}$ satisfies

$$\mathbb{E}[N_{\delta,\bar{\mu}}] \geq \frac{\log(1/2.4\delta)}{D^*(\mu_1, \mu_2)}, \tag{3}$$

where $D^*(x, y) \coloneqq d\left(x, \frac{x+y}{2}\right) + d\left(y, \frac{x+y}{2}\right)$.

The proof of Corollary 2 is provided in Appendix B. Our upper bounds in Section IV for Bernoulli bandits asymptotically match the lower bound from the corollary. While it may

be surprising that the lower bound depends only on the separation between the top two arms, recall that under uniform sampling, each round produces a sample from every arm. Our experiments in Section V do suggest a slow growth in round complexity as more sub-optimal arms are added while keeping the top two fixed; analytically this growth comes from terms that are sub-linear in $\log(\frac{1}{\delta})$.

IV. δ-PAC Stopping Rules for Uniform Sampling

A key technical challenge in the design of algorithms in the fixed confidence setting is to deal with *random* stopping time, which is necessary to be efficient on easy problem instances. The recent development of "anytime" confidence bounds in the literature addresses this issue. We develop stopping rules for the uniform sampling setting based on two such approaches. We propose these rules and analyse them in the context of Bernoulli bandits. However, a standard procedure [26, See Section 1.2] can generate consistent Bernoulli samples from any reward distribution with a known, bounded support—so our upper bounds apply to this wider range.

A. Prior-Posterior Ratio Martingale Based Stopping Rule

Consider any family of distributions $\{\Pi_p\}_{p\in\mathcal{P}}$ parameterised by p with the density function $\pi_p(x)$. Suppose there is a "ground truth" parameter $p^* \in \mathcal{P}$, which we wish to estimate. In the Bayesian approach, we take some initial prior over $\mathcal{P}$, say $f_0(p)$. Now, after collecting t samples $X \equiv (X_1, X_2, \ldots, X_t)$ from the distribution Π_{p^*}, the posterior $f_t(p)$ is given by:

$$f_t(p) = \frac{f_0(p)\mathcal{L}_p(X)}{\int_{q\in\mathcal{P}} f_0(q)\mathcal{L}_q(X)dq},$$

where $\mathcal{L}_p(X)$ $(= \prod_{i=1}^t \pi_p(X_i)$ if samples are i.i.d.) gives the likelihood of the outcomes for a given parameter p. Then, the prior-posterior ratio (PPR) at time t is the quantity

$$R_t(p) \coloneqq \frac{f_0(p)}{f_t(p)}.$$

Waudby-Smith and Ramdas [24, see Proposition 2.1] define

$$C_t \coloneqq \left\{p \in \mathcal{P} \mid R_t(p) < \frac{1}{\delta}\right\},$$

and show that $(C_t)_{t=0}^\infty$ is a *confidence sequence*, as below.

Proposition 3 (PPR Martingale): For any prior $f_0(p)$ on $\mathcal{P}$ that assigns non-zero mass everywhere, the sequence of prior-posterior ratios evaluated at the true parameter p^*, that is $(R_t(p))_{t=0}^\infty$, is a non-negative martingale with respect to $(\mathcal{F}_t = \sigma(X))_{t=0}^\infty$. Further, the sequence of sets C_t forms a $(1-\delta)$ confidence sequence for p^*: that is,

$$\mathbb{P}(\exists t \geq 0 : p^* \notin C_t) \leq \delta \iff \mathbb{P}(\forall t \geq 0 : p^* \in C_t) \geq 1-\delta.$$

In other words, the true parameter p^* remains inside the confidence sequence for all time t with probability at least $1-\delta$. This result generalises to the estimation of multiple parameters [24]. Following a similar approach as the application of this idea to estimate the mode of a discrete distribution [25], we work out a rule for BAI with uniform sampling.

1) $K = 2$ Bernoulli Arms: Consider the case where we have only $K = 2$ Bernoulli arms with means μ_1 and μ_2. We will try to jointly estimate the two parameters (μ_1, μ_2). Since, Bernoulli random variables work with the Beta distribution as a conjugate prior, and we need a prior with non-zero mass everywhere, a uniform prior is a suitable choice. Therefore, we have the prior given by $f_0(p_1, p_2) = 1$ for $p_1, p_2 \in [0, 1]$.

After n rounds, suppose arm 1 has yielded s_1^n 1's and f_1^n 0's, while arm 2 has yielded s_2^n 1's and f_2^n 0's. Note that, $s_1^n + f_1^n = s_2^n + f_2^n = n$. Since the reward distributions are independent and Bernoulli, we have,

$$f_n(p_1, p_2) = \text{Beta}(p_1;\, s_1^n+1, f_1^n+1)\,\text{Beta}(p_2;\, s_2^n+1, f_2^n+1).$$

The corresponding $(1-\delta)$ confidence sequence becomes

$$C_n = \left\{(p_1, p_2) \in [0,1]^2 \mid f_n(p_1, p_2) > \delta\right\}.$$

Now, we aim to determine which arm dominates the other. So, it suffices for us to stop when C_n only contains points (p_1, p_2) such that (1) *all* points satisfy $p_1 > p_2$, in which case 1 is the winner, or (2) *all* points satisfy $p_2 > p_1$, making 2 the winner. Without loss of generality, assume that 1 is the empirically superior class: that is, $s_1^n > s_2^n$ (if $s_1^n = s_2^n$, a reasonable algorithm will not stop). Hence, at termination, C_n will necessarily have (p_1, p_2) pairs in which $p_1 > p_2$. In general, could C_n also have pairs in which $p_2 \geq p_1$? The following lemma (proven in Appendix C) provides an easy way to verify.

Lemma 4: Consider $(p_1, p_2) \in [0, 1]^2$ such that $p_1 < p_2$, and let $\bar{p} \coloneqq \frac{p_1+p_2}{2}$. If $s_1^n > s_2^n$, then $f_n(p_1, p_2) < f_n(\bar{p}, \bar{p})$.

The lemma implies that if 1 is the empirically superior class, then C_n can contain "bad" points, (p_1, p_2) with $p_1 \leq p_2$ only if it also contains some point of the form (p, p). Consequently, for us to stop, it suffices for C_n to separate from the line $p_1 = p_2$. Our algorithm can be simplified to stop as soon as C_n no longer contains any point of the form (p, p), and then declaring the empirically superior arm as the winner. We make another useful observation in this regard. For $p \in [0, 1]$:

$$\begin{aligned} f_n(p,p) &= \text{Beta}(p; s_1^n+1, f_1^n+1)\text{Beta}(p; s_2^n+1, f_2^n+1) \\ &= \frac{p^{s_1^n}(1-p)^{f_1^n} p^{s_2^n}(1-p)^{f_2^n}}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &= \frac{B(s_1^n+s_2^n+1, f_1^n+f_2^n+1)}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &\quad \times \text{Beta}(p; s_1^n+s_2^n+1, f_1^n+f_2^n+1). \end{aligned}$$

In other words, $f_n(p, p)$ can be represented as a *single* Beta pdf, with some multiplicative factors independent of p. The mode of this Beta pdf occurs at $\hat{\mu}_{1,2} \coloneqq \frac{s_1^n+s_2^n}{2n} = \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$. Hence, our stopping rule can effectively be stated as: stop as soon as $f_n(\hat{\mu}_{1,2}, \hat{\mu}_{1,2}) \leq \delta$, that is:

$$\begin{aligned} \ln(s_1^n!) + \ln(s_2^n!) + \ln(f_1^n!) + \ln(f_2^n!) - 2\ln((n+1)!) \\ -(s_1^n+s_2^n)\ln(s_1^n+s_2^n) - (f_1^n+f_2^n)\ln(f_1^n+f_2^n) \\ +2n\ln(2n) \geq \ln\left(\frac{1}{\delta}\right). \quad (4) \end{aligned}$$

Since we used the *joint distribution* over two arms for generating our confidence sequence, we call this rule PPR-JD. This stopping rule needs only a constant number of arithmetic and logarithmic operations on each round, since the "log factorial" terms for round $n+1$ are computable in constant time if the corresponding terms for round n are stored.

2) $K \geq 2$ Bernoulli arms: The PPR-JD stopping rule can be easily extended to instances with $K \geq 2$ arms. Like the "1 versus 1" approach of Jain et al. [25], we consider all pairs of arms, and check if there exists an arm that is *empirically better* than every other arm according to the PPR-JD rule for two arms. If such an arm exists, clearly it must be α_1 (the empirical best). Consequently it suffices to check the PPR-JD rule only with the $K-1$ pairs that include α_1. To ensure that the overall mistake probability does not exceed δ, each pair must follow the PPR-JD rule with $\delta_K = \delta/(K-1)$. In summary, our rule could be to stop if and only if $f_n(\hat{\mu}_{\alpha_1,b}, \hat{\mu}_{\alpha_1,b}) \leq \frac{\delta}{K-1}$ for $b \in [K]\setminus\{\alpha_1\}$. Interestingly, a working in Appendix D shows that $f_n(\hat{\mu}_{\alpha_1,\alpha_2}, \hat{\mu}_{\alpha_1,\alpha_2}) \geq f_n(\hat{\mu}_{\alpha_1,b}, \hat{\mu}_{\alpha_1,b})\ \forall\ b \in [K] \setminus \{\alpha_1, \alpha_2\}$, where α_2 is the arm with the second highest empirical mean. Hence, the PPR-JD rule further simplifies to comparing only the top two empirically best arms.

Stop and return α_1 iff $f_n(\hat{\mu}_{\alpha_1,\alpha_2}, \hat{\mu}_{\alpha_1,\alpha_2}) \leq \dfrac{\delta}{K-1}$.

B. Chernoff's Stopping Rule for Uniform Sampling

We now describe the δ-PAC guarantee obtained using Chernoff's stopping rule along with an informational threshold, as outlined in by Garivier and Kaufmann [17]. We adapt their approach to uniform sampling.

The Generalized Likelihood Ratio statistic for arms $a, b \in [K]$ gives the log ratio of the maximum likelihood of arm a having higher mean than arm b, over the opposite hypothesis:

$$\Lambda_{a,b}(t) := \log\left(\frac{\max_{\mu'_a \geq \mu'_b} \mathcal{L}_{\mu'_a}(X_a)\mathcal{L}_{\mu'_b}(X_b)}{\max_{\mu'_a \leq \mu'_b} \mathcal{L}_{\mu'_a}(X_a)\mathcal{L}_{\mu'_b}(X_b)}\right).$$

Intuitively, a higher $\Lambda_{a,b}(t)$ value places a higher belief on arm a being better than arm b. It is well known that this term has an analytical closed-form representation for Bernoulli Bandits [17]. After adapting it to uniform sampling, we get:

$$\begin{aligned}\Lambda_{a,b}(n) &= n\left[d\left(\hat{\mu}_a(n), \frac{\hat{\mu}_a(n)+\hat{\mu}_b(n)}{2}\right)\right. \\ &\qquad \left. + d\left(\hat{\mu}_b(n), \frac{\hat{\mu}_a(n)+\hat{\mu}_b(n)}{2}\right)\right] \\ &= nD^*(\hat{\mu}_a(n), \hat{\mu}_b(n)) \end{aligned} \tag{5}$$

if $\hat{\mu}_a \geq \hat{\mu}_b$. By definition it follows that $\Lambda_{a,b}(n) = -\Lambda_{b,a}(n)$. Garivier and Kaufmann [17] suggest an intuitive "1 versus 1" stopping rule based on this statistic exceeding a suitable threshold $\beta(n,\delta)$. They propose stopping if and only if

$$\max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} \Lambda_{a,b}(n) > \beta(n,\delta).$$

Since $\Lambda_{a,b}(n)$ is only non-negative when arm a is empirically superior to arm b, the outer maximiser is clearly α_1, the arm with the highest empirical mean. Combining this with the fact that $D^*(x,y)$ is decreasing in y for $x > y$, we obtain:

$$\begin{aligned}\Lambda(n) &= \max_{a\in[K]} \min_{b\in[K]\setminus\{a\}} \Lambda_{a,b}(n) = \min_{b\in[K]\setminus\{\alpha_1\}} \Lambda_{\alpha_1,b}(n) \\ &= nD^*(\hat{\mu}_{\alpha_1}(n), \hat{\mu}_{\alpha_2}(n)).\end{aligned}$$

To provide a δ-PAC guarantee, we must appropriately set the threshold $\beta(n,\delta)$. To this end, Garivier and Kaufmann [17, see Theorem 10] provide an informational threshold for Bernoulli bandits, which they prove to be δ-PAC for any sampling strategy. We can directly use their choice, to set

$$\beta(n,\delta) = \log\left(\frac{2nK(K-1)}{\delta}\right).$$

The resulting stopping rule, which we denote U-CNF (for Uniform-Chernoff), is as given below.

Stop and return α_1 iff $nD^*(\hat{\mu}_{\alpha_1}(n), \hat{\mu}_{\alpha_2}(n)) > \beta(n,\delta)$.

Observe that both stopping rules only depend on the top two empirically superior arms. This aspect is distinctive to the setting of uniform sampling, and not generally true [17]. As we see next, both rules are asymptotically optimal as $\delta \to 0$.

C. Asymptotic Optimality of the Stopping Rules

The following theorem, adapted from Garivier and Kaufmann [17, see Theorem 14], is central to our analysis. Although PPR-JD and U-CNF are different rules, it is on account of satisfying the conditions of this theorem that they become asymptotically optimal.

Theorem 5 (Asymptotic Upper Bound): Let $\bar{\mu}$ be any K-armed Bernoulli bandit instance. Then, under uniform sampling, any rule that stops and returns α_1 if and only if

$$\Lambda(n) > \log\left(\frac{Cn}{\delta}\right)$$

for some positive constant C, satisfies the following upper bound on its expected stopping time:

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1}{D^*(\mu_1,\mu_2)}.$$

We obtain the following corollaries for PPR-JD and U-CNF.

Corollary 6 (Asymptotic Optimality of PPR-JD): Let $\bar{\mu}$ be any K-armed Bernoulli bandit instance. Then, the stopping rule PPR-JD satisfies the following upper bound on its expected stopping time:

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1}{D^*(\mu_1,\mu_2)}.$$

Corollary 7 (Asymptotic Optimality of U-CNF): Let $\bar{\mu}$ be any K-armed Bernoulli bandit instance. Then, the stopping rule U-CNF satisfies the following upper bound on its expected stopping time:

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1}{D^*(\mu_1,\mu_2)}.$$

Detailed proofs of Theorem 5 and Corollary 6 are given in appendices E and F, respectively. Corollary 7 follows directly from Theorem 5 by taking $C = 2K(K-1)$.

V. Empirical Results

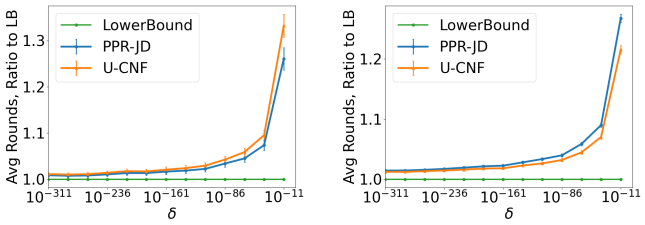
In this section, we present an empirical evaluation and comparison of PPR-JD and U-CNF. To begin, we confirm that both algorithms are indeed asymptotically optimal as $\delta \to 0$. Figure 1 shows the ratio of the round complexity of either method to the lower bound in Corollary 2 on two different Bernoulli bandit instances, both 2-armed. The round complexity is the average of 200 independent runs; the plots also show one standard error. The x axis in plots (a) and (b) shows mistake probability δ varied in the range $[10^{-311}, 10^{-11}]$; notice that the ratio approaches 1 for both methods at the lower end of this spectrum.

For any fixed value of δ, which among our two methods is more sample-efficient? Interestingly, we observe that on "typical" instances, in which the arms' means are well-separated from 0 and 1, PPR-JD terminates well before U-CNF (see, for example, Figure 1a). However, when the means are close to 0 and 1 (see, for example, Figure 1b), U-CNF enjoys a slight advantage. Note that the latter instance is actually an easy instance for BAI, since the arms are well-separated. The empirical trend noted here is predicted by Lemma 9 in Appendix F. In particular, the "threshold" we obtain for PPR-JD contains the quantity

$$h_{\alpha_1,\alpha_2}(n) = \sqrt{\hat{\mu}_{\alpha_1}(1-\hat{\mu}_{\alpha_1})\hat{\mu}_{\alpha_2}(1-\hat{\mu}_{\alpha_2})}$$

in the denominator (within the log). Clearly h_{α_1,α_2} will be small when the top two means in the instance get close to the extremes. Still, we find U-CNF to dominate only moderately, that too only for means as close to 1 as 0.97 or close to 0 as 0.01. For most realistic bandit instances, PPR-JD would appear to be the method of choice.

Table I illustrates the performance of our stopping rules as the number of arms K is increased. Once again we report the average number of rounds before stopping divided by the lower bound from Corollary 2. As expected, notice that both algorithms have a ratio quite close to 1 for the lower value of $\delta = 10^{-261}$. The round complexity is largely determined by the top two arms. As more arms are added (see rows 2–4), there is an increase in the sample complexity (note that the δ-dependent lower bound remains unaffected). The increase due to more arms is gradual when the means of the arms are well separated from the second best (rows 2–4), and more pronounced when all sub-optimal arms are given the same mean (rows 5–7). Nonetheless, these additional arms do not affect the asymptotic sample complexity (as $\delta \to 0$).



(a) $\mu_1 = 0.7, \mu_2 = 0.5$ (b) $\mu_1 = 0.99, \mu_2 = 0.01$

Fig. 1: Plots on two 2-armed Bernoulli bandit instances showing the ratio of the average number of rounds to the lower bound as δ is varied.

TABLE I

Stopping times for $K \geq 2$ Bernoulli arms. Results are averages from 200 independent runs, and show one standard error.

Bandit Instance ($\bar{\mu}$)	δ	**Average rounds / Lower bound**	
		PPR-JD	U-CNF
(0.7, 0.3)	10^{-11}	1.215 ± 0.022	1.277 ± 0.023
	10^{-261}	1.007 ± 0.004	1.010 ± 0.004
(0.7, 0.3, 0.24)	10^{-11}	1.257 ± 0.021	1.336 ± 0.022
	10^{-261}	1.013 ± 0.004	1.016 ± 0.004
(0.7, 0.3, 0.24, 0.18)	10^{-11}	1.284 ± 0.021	1.378 ± 0.022
	10^{-261}	1.006 ± 0.004	1.010 ± 0.004
(0.7, 0.3, 0.24, 0.18, 0.12)	10^{-11}	1.286 ± 0.021	1.390 ± 0.022
	10^{-261}	1.011 ± 0.004	1.015 ± 0.004
(0.7, 0.3, 0.3)	10^{-11}	1.370 ± 0.022	1.455 ± 0.023
	10^{-261}	1.033 ± 0.004	1.036 ± 0.004
(0.7, 0.3, 0.3, 0.3)	10^{-11}	1.472 ± 0.024	1.565 ± 0.024
	10^{-261}	1.044 ± 0.004	1.047 ± 0.004
(0.7, 0.3, 0.3, 0.3, 0.3)	10^{-11}	1.503 ± 0.022	1.627 ± 0.023
	10^{-261}	1.052 ± 0.003	1.056 ± 0.003

In summary, these experiments validate our analytical findings about PPR-JD and U-CNF, and suggest PPR-JD is preferable except on instances with means very close to 0 or 1.

VI. Conclusion

We have considered the problem of BAI in the fixed confidence setting, when all the arms get sampled in each round. Our *uniform sampling* setting is relevant in applications wherein arms can be sampled in parallel. Since only stopping (rather than the choice of sampling) depends on the outcomes of pulls in this setting, it is less affected by practical violations such as delay in obtaining rewards [27,28].

From a theoretical standpoint, we provide a complete characterisation of problem and solution. First, we suitably adapt the lower bound for Kaufmann et al. [23] for uniform sampling. Next, we propose two separate stopping rules, PPR-JD and U-CNF, based on existing algorithmic frameworks in the literature. Both rules are simpler and more computationally efficient than "fully sequential" counterparts [17]–[19]. By adapting existing analyses [17,25], we obtain instance-specific upper bounds for both rules. We observe that both rules are asymptotically optimal with respect to the lower bound for Bernoulli bandits. However, as apparent from Figure 1, there remains a gap between upper and lower bounds for finite (non-asymptotic) values of δ. Future work could pursue a supplementary lower bound possibly showing a dependence on K, albeit as a coefficient to a sub-$\log(\frac{1}{\delta})$ factor and extend these methods to other reward distributions.

Our analytical results are reaffirmed by experiments, which also provide guidance for choosing between PPR-JD and U-CNF in practice. The conceptual simplicity of uniform sampling, coupled with the ease of implementing PPR-JD and U-CNF, is likely to benefit practitioners.

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2–3, pp. 235–256, 2002.

[2] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7568. Springer, 2012, pp. 199–213.

[3] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, 2006.

[4] B. Thananjeyan, K. Kandasamy, I. Stoica, M. I. Jordan, K. Goldberg, and J. E. Gonzalez, "PAC best arm identification under a deadline," *CoRR*, vol. abs/2106.03221, 2021.

[5] F. Yao, C. Li, D. Nekipelov, H. Wang, and H. Xu, "Learning the optimal recommendation from explorative users," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 9457–9465.

[6] Y. Didwania, J. Nair, and N. Hemachandra, "Unsupervised crowdsourcing with accuracy and cost guarantees," in *20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks, WiOpt 2022, Torino, Italy, September 19-23, 2022*. IEEE, 2022, pp. 137–144.

[7] Y. Zhou, X. Chen, and J. Li, "Optimal pac multiple arm identification with applications to crowdsourcing," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 217–225.

[8] X. Zhou, B. Hao, J. Kang, T. Lattimore, and L. Li, "Sequential best-arm identification with application to brain-computer interface," *CoRR*, vol. abs/2305.11908, 2023.

[9] E. Kaufmann and W. M. Koolen, "Monte-carlo tree search by best arm identification," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 4897–4906.

[10] J. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*. Omnipress, 2010, pp. 41–53.

[11] Z. S. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 1238–1246.

[12] A. Carpentier and A. Locatelli, "Tight (lower) bounds for the fixed budget best arm identification bandit problem," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 49. JMLR.org, 2016, pp. 590–604.

[13] C. Qin, "Open problem: Optimal best arm identification with fixed-budget," in *Proceedings of Thirty Fifth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 178. PMLR, 02–05 Jul 2022, pp. 5650–5654.

[14] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proceedings of the Twenty-ninth International Conference on Machine Learning (ICML 2012)*. New York, NY, USA: Omnipress, 2012, pp. 655–662.

[15] E. Kaufmann and S. Kalyanakrishnan, "Information complexity in bandit subset selection," in *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, ser. JMLR Workshop and Conference Proceedings, vol. 30. JMLR.org, 2013, pp. 228–251.

[16] K. G. Jamieson, M. Malloy, R. D. Nowak, and S. Bubeck, "lil' UCB : An optimal exploration algorithm for multi-armed bandits," in *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, ser. JMLR Workshop and Conference Proceedings, vol. 35. JMLR.org, 2014, pp. 423–439.

[17] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 49. JMLR.org, 2016, pp. 998–1027.

[18] D. Russo, "Simple bayesian algorithms for best arm identification," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 49. JMLR.org, 2016, pp. 1417–1418.

[19] X. Shang, R. de Heide, P. Menard, E. Kaufmann, and M. Valko, "Fixed-confidence guarantees for bayesian best-arm identification," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 26–28 Aug 2020, pp. 1823–1832.

[20] A. Atsidakou, S. Katariya, S. Sanghavi, and B. Kveton, "Bayesian fixed-budget best-arm identification," *CoRR*, vol. abs/2211.08572, 2022.

[21] D. Urieli, P. MacAlpine, S. Kalyanakrishnan, Y. Bentor, and P. Stone, "On optimizing interdependent skills: A case study in simulated 3d humanoid robot soccer," in *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, vol. 2. IFAAMAS, 2011, pp. 769–776.

[22] C. Payne, "Election forecasting in the UK: The BBC's experience," *Euramerica*, vol. 33, no. 1, pp. 193–234, 2003.

[23] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *J. Mach. Learn. Res.*, vol. 17, pp. 1:1–1:42, 2016.

[24] I. Waudby-Smith and A. Ramdas, "Confidence sequences for sampling without replacement," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[25] S. Anand Jain, R. Shah, S. Gupta, D. Mehta, I. J. Nair, J. Vora, S. Khyalia, S. Das, V. J. Ribeiro, and S. Kalyanakrishnan, "Pac mode estimation using ppr martingale confidence sequences," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 151. PMLR, 28–30 Mar 2022, pp. 5815–5852.

[26] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Mannor, N. Srebro, and R. C. Williamson, Eds., vol. 23. Edinburgh, Scotland: PMLR, 25–27 Jun 2012, pp. 39.1–39.26. [Online]. Available: https://proceedings.mlr.press/v23/agrawal12.html

[27] W. Tang, C.-J. Ho, and Y. Liu, "Bandit learning with delayed impact of actions," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 26 804–26 817.

[28] C. Pike-Burke, S. Agrawal, C. Szepesvari, and S. Grunewalder, "Bandits with delayed, aggregated anonymous feedback," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 4105–4113.

[29] F. Topsoe, "Some bounds for the logarithmic function," *Inequality Theory and Applications*, vol. 4, 01 2007.

[30] W. Mulzer, "Five proofs of chernoff's bound with applications," *CoRR*, vol. abs/1801.03365, 2018.

[31] H. Robbins, "A remark on stirling's formula," *The American Mathematical Monthly*, vol. 62, no. 1, pp. 26–29, 1955.

APPENDIX A
PROOF OF THEOREM 1

Proof: This proof borrows from Garivier and Kaufmann [17], who, in turn, adapt the lower bound proof given by Kaufmann et al. [23]. From [17, Proof of Theorem 1], we have for $\delta \in (0,1)$ and any valid sampling and stopping rule:

$$d(\delta, 1-\delta) \leq \inf_{\mathcal{I}' \in \Omega'(\mathcal{I})} \left(\sum_{a=1}^{K} \mathbb{E}\left[N^a_{\delta,\mathcal{I}}\right] \mathrm{KL}(\Pi_a \parallel \Pi'_a) \right)$$

where $N^a_{\delta,\mathcal{I}}$ denotes the number of pulls of arm a at stopping. Since, under uniform sampling, $\mathbb{E}[N^a_{\delta,\mathcal{I}}] = \mathbb{E}[N_{\delta,\mathcal{I}}]$, we have:

$$d(\delta, 1-\delta) \leq \mathbb{E}[N_{\delta,\mathcal{I}}] \inf_{\mathcal{I}' \in \Omega'(\mathcal{I})} \left(\sum_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a) \right).$$

The above statement can be simplified using the easily verifiable inequality $d(\delta, 1-\delta) \geq \log(1/2.4\delta)$ leading to,

$$\mathbb{E}[N_{\delta,\mathcal{I}}] \geq \frac{\log(1/2.4\delta)}{\inf\limits_{\mathcal{I}' \in \Omega'(\mathcal{I})} \sum\limits_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a)}$$

■

APPENDIX B
PROOF OF COROLLARY 2

Proof: We have

$$\Omega'(\mathcal{I}) = \{\mathcal{I}' \mid a^*(\mathcal{I}') \neq a^*(\mathcal{I}),\, \mathcal{I}' \in \Omega\}$$

which for $\bar{\mu} \in \mathcal{B}$ can be rewritten as

$$\begin{aligned}\Omega'(\bar{\mu}) &= \{\bar{\mu}' \mid a^*(\bar{\mu}') \neq 1,\, \bar{\mu}' \in \Omega\} \\ &= \bigcup_{a \neq 1} \{\bar{\mu}' \mid \mu'_a > \mu'_1,\, \bar{\mu}' \in \Omega\}.\end{aligned}$$

Then,

$$\begin{aligned}\inf_{\mathcal{I}' \in \Omega'(\mathcal{I})} \sum_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a) &= \inf_{\bar{\mu}' \in \Omega'(\bar{\mu})} \sum_{a=1}^{K} d(\mu_a, \mu'_a) \\ &= \min_{a \neq 1} \inf_{\bar{\mu}' : \mu'_a > \mu'_1} \sum_{a=1}^{K} d(\mu_a, \mu'_a) \\ &= \min_{a \neq 1} \inf_{\bar{\mu}' : \mu'_a \geq \mu'_1} d(\mu_1, \mu'_1) \\ &\qquad + d(\mu_a, \mu'_a)\end{aligned}$$

where the last equality comes from the fact that, for the inner infimum, we can let $\mu'_b = \mu_b \,\forall\, b \in [K] \setminus \{1, a\}$. Now, the inner infimum is a constrained convex optimization problem since $d(x,y)$ is convex. It can be shown that it has an analytical solution given by $D^*(\mu_1, \mu_a)$. Therefore,

$$\inf_{\mathcal{I}' \in \Omega'(\mathcal{I})} \sum_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a) = \min_{a \neq 1} D^*(\mu_1, \mu_a).$$

It can be easily verified that $D^*(x,y)$ is decreasing in y for $x > y$. Therefore,

$$\inf_{\mathcal{I}' \in \Omega'(\mathcal{I})} \sum_{a=1}^{K} \mathrm{KL}(\Pi_a \parallel \Pi'_a) = D^*(\mu_1, \mu_2).$$

The corollary then follows from plugging the above relation in the statement of Theorem 1. ■

APPENDIX C
PROOF OF LEMMA 4

Define $\Delta := \bar{p} - p_1 = p_2 - \bar{p}$. We have,

$$\begin{aligned}&f_n(p_1, p_2) \\ &= \text{Beta}(p_1; s_1^n + 1, f_1^n + 1)\text{Beta}(p_2; s_2^n + 1, f_2^n + 1) \\ &= \frac{(p_1)^{s_1^n}(1-p_1)^{f_1^n}(p_2)^{s_2^n}(1-p_2)^{f_2^n}}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &= (p_1 p_2)^{s_2^n}\left((1-p_1)(1-p_2)\right)^{f_1^n} \\ &\quad \times \frac{(p_1)^{s_1^n - s_2^n}(1-p_2)^{f_2^n - f_1^n}}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &= (\bar{p}^2 - \Delta^2)^{s_2^n}\left((1-\bar{p})^2 - \Delta^2\right)^{f_1^n} \\ &\quad \times \frac{(\bar{p} - \Delta)^{s_1^n - s_2^n}(1-\bar{p} - \Delta)^{f_2^n - f_1^n}}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &< \frac{(\bar{p}^2)^{s_2^n}\left((1-\bar{p})^2\right)^{f_1^n}(\bar{p})^{s_1^n - s_2^n}(1-\bar{p})^{f_2^n - f_1^n}}{B(s_1^n+1, f_1^n+1)B(s_2^n+1, f_2^n+1)} \\ &= \text{Beta}(\bar{p}; s_1^n + 1, f_1^n + 1)\text{Beta}(\bar{p}; s_2^n + 1, f_2^n + 1) \\ &= f_n(\bar{p}, \bar{p}).\end{aligned}$$

APPENDIX D
SUFFICIENCY OF TOP TWO ARMS FOR PPR-JD

In this section, we prove that implementing the PPR-JD rule for K arms is equivalent to considering the PPR-JD rule for just the top two arms in terms of empirical mean rewards. We show this by arguing that,

$$f_n(\hat{\mu}_{\alpha_1,\alpha_2}, \hat{\mu}_{\alpha_1,\alpha_2}) \geq f_n(\hat{\mu}_{\alpha_1,b}, \hat{\mu}_{\alpha_1,b}) \,\forall\, b \in [K] \setminus \{\alpha_1, \alpha_2\}. \tag{6}$$

For ease of notation we drop the α's and use numeric indices. We prove the inequality (6) by showing that for any $\hat{\mu}_b$ such that, $\hat{\mu}_1 > \hat{\mu}_2 \geq \hat{\mu}_b$, we have that,

$$f_n(\hat{\mu}_{1,2}, \hat{\mu}_{1,2}) \geq f_n(\hat{\mu}_{1,b}, \hat{\mu}_{1,b}).$$

We can restate this in terms of the number of successes $s_a = n\hat{\mu}_a$. Then, for $g(s_a, s_b) \coloneqq f_n(\hat{\mu}_{a,b}, \hat{\mu}_{a,b})$, we have to show that:

$$g(s_1, s_2) \geq g(s_1, s_b) \,\forall\, s_b \leq s_2 < s_1. \tag{7}$$

If we can show that

$$g(s_1, s_2) \geq g(s_1, s_2 - 1) \,\forall\, s_1 > s_2 \geq 1 \tag{8}$$

then (7) easily follows since

$$g(s_1, s_2) \geq g(s_1, s_2 - 1) \geq \ldots \geq g(s_1, s_b).$$

Therefore, we just need to prove inequality (8).

By expanding the inequality (8), we can simplify our objective to showing that $\forall\, n \geq s_1 > s_2 \geq 1$:

$$\left(1 - \frac{n - s_1}{2n + 1 - s_1 - s_2}\right)\left(1 + \frac{s_1}{s_2}\right)$$
$$\left(1 + \frac{1}{s_1 + s_2 - 1}\right)^{s_1 + s_2 - 1}$$
$$\left(1 - \frac{1}{2n + 1 - s_1 - s_2}\right)^{2n - s_1 - s_2} \geq 1.$$

Now, we can easily check that the first two terms are decreasing and the last two terms are increasing in s_2. Therefore by substituting s_2 with $s_1 - 1$ for the first two terms and by 1 for the last two terms, we have that:

$$\left(1 - \frac{n - s_1}{2n + 1 - s_1 - s_2}\right)\left(1 + \frac{s_1}{s_2}\right)$$
$$\left(1 + \frac{1}{s_1 + s_2 - 1}\right)^{s_1 + s_2 - 1}$$
$$\left(1 - \frac{1}{2n + 1 - s_1 - s_2}\right)^{2n - s_1 - s_2}$$
$$\geq \left(1 - \frac{n - s_1}{2n + 2 - 2s_1}\right)\left(1 + \frac{s_1}{s_1 - 1}\right)$$
$$\left(1 + \frac{1}{s_1}\right)^{s_1}\left(1 - \frac{1}{2n - s_1}\right)^{2n - s_1 - 1}$$
$$= \left(\frac{n + 2 - s_1}{2(n + 1 - s_1)}\right)\left(2 + \frac{1}{s_1 - 1}\right)$$
$$\left(1 + \frac{1}{s_1}\right)^{s_1}\left(1 - \frac{1}{2n - s_1}\right)^{2n - s_1 - 1}.$$

Now, the first term is $\geq 1/2$ and the last term is $\geq 1/e$. Therefore,

$$\left(\frac{n + 2 - s_1}{2(n + 1 - s_1)}\right)\left(2 + \frac{1}{s_1 - 1}\right)$$
$$\left(1 + \frac{1}{s_1}\right)^{s_1}\left(1 - \frac{1}{2n - s_1}\right)^{2n - s_1 - 1}$$
$$\geq \frac{1}{2e}\left(2 + \frac{1}{s_1 - 1}\right)\left(1 + \frac{1}{s_1}\right)^{s_1}$$
$$\geq \frac{1}{2e}\left(2 + \frac{1}{s_1}\right)\left(1 + \frac{1}{s_1}\right)^{s_1}$$
$$= \frac{1}{2e}\left(\left(1 + \frac{1}{s_1}\right)^{s_1} + \left(1 + \frac{1}{s_1}\right)^{s_1 + 1}\right).$$

Now, if we can show that $y(s_1) := (1 + 1/s_1)^{s_1} + (1 + 1/s_1)^{s_1 + 1} \geq 2e$, then we are done. We show that the function $y(x)$ is decreasing for $x > 0$. We differentiate and get that,

$$y'(x) = \left(\frac{1}{x} + 1\right)^x \left(\ln\left(\frac{1}{x} + 1\right) - \frac{1}{x + 1}\right)$$
$$+ \left(\frac{1}{x} + 1\right)^{x + 1}\left(\ln\left(\frac{1}{x} + 1\right) - \frac{1}{x}\right)$$
$$= \left(\frac{1}{x} + 1\right)^x \left(\left(2 + \frac{1}{x}\right)\ln\left(\frac{1}{x} + 1\right) - \frac{2x^2 + 2x + 1}{x^2(x + 1)}\right).$$

Now, the first term is positive. We will next show that the second term is negative. Using the upper bound of inequality (3) in [29], we have that:

$$\left(2 + \frac{1}{x}\right)\ln\left(\frac{1}{x} + 1\right) - \frac{2x^2 + 2x + 1}{x^2(x + 1)}$$
$$\leq \left(2 + \frac{1}{x}\right)\frac{1}{2x}\frac{2x + 1}{x + 1} - \frac{2x^2 + 2x + 1}{x^2(x + 1)}$$
$$= \frac{2x^2 + 2x + 1/2}{x^2(x + 1)} - \frac{2x^2 + 2x + 1}{x^2(x + 1)}$$
$$= -\frac{1}{2x^2(x + 1)} < 0.$$

Hence, we have shown that $y'(x) < 0$, and therefore $y(x)$ is decreasing. It is easy to check that $y(x)$ has the limit $2e$ for $x \to \infty$. Therefore, we have that, $y(s_1) \geq 2e$, as required.

Appendix E
Proof of Theorem 5

Proof: Our proof proceeds along similar lines as the proof of Theorem 14 in [17]. We begin by considering the event that the empirical reward averages for all the arms are close to their actual means, which is highly probable due to the law of large numbers. Assuming this to be true, we then give a bound on the stopping time under this condition. We then bound the probability of the case where the arm's empirical reward averages are not close to their actual means and show that the expected stopping time satisfies the upper bound in the theorem statement, as $\delta \to 0$.

We define the event $\mathcal{E}_n(\eta)$ as the event that the empirical reward averages for all the arms are in η-neighbourhood of their actual means at round n. Formally,

$$\mathcal{E}_n(\eta) := \Big\{|\hat{\mu}_a(n) - \mu_a| < \eta\ \forall\, a \in [K]\Big\}.$$

Now we can choose η' small enough ($\eta' < \min_{a,b \in [K]} |\mu_a - \mu_b|/2$) such that there is no overlap between the neighbourhoods around each mean. Then under the event $\mathcal{E}_n(\eta')$, we have $\hat{\mu}_1(n) > \hat{\mu}_2(n) \geq \hat{\mu}_3(n) \cdots \geq \hat{\mu}_K(n)$. Therefore, $\alpha_1 = 1, \alpha_2 = 2, \ldots \alpha_K = K$. From here on, for brevity, we abuse the notation $\hat{\mu}_a = \hat{\mu}_a(n)$.

Continuing under this event $\mathcal{E}_n(\eta')$ we would have at round n that,

$$\Lambda(n) = nD^*(\hat{\mu}_{\alpha_1}(n), \hat{\mu}_{\alpha_2}(n))$$
$$= nD^*(\hat{\mu}_1, \hat{\mu}_2)$$
$$\geq \inf_{\substack{|\tilde{\mu}_a - \mu_a| < \eta' \\ a = 1,2}} nD^*(\tilde{\mu}_1, \tilde{\mu}_2).$$

Now for any $\epsilon > 0$, we can choose $\eta'' := \eta(\epsilon)$ small enough such that,

$$\inf_{\substack{|\tilde{\mu}_a - \mu_a| < \eta'' \\ a = 1,2}} D^*(\tilde{\mu}_1, \tilde{\mu}_2) \geq \frac{D^*(\mu_1, \mu_2)}{1 + \epsilon}.$$

Therefore, we have for $\eta^* := \min(\eta', \eta'')$ that under the event $\mathcal{E}_n(\eta^*)$, at round n,

$$\Lambda(n) \geq \frac{nD^*(\mu_1, \mu_2)}{1+\epsilon}.$$

We define the quantity $N^* \equiv N^*(\delta, \epsilon, \bar{\mu})$,

$$N^* := \frac{1+\epsilon}{D^*(\mu_1,\mu_2)} \left[\log\left(\frac{1}{\delta} \cdot \frac{Ce(1+\epsilon)}{D^*(\mu_1,\mu_2)}\right) + \log\log\left(\frac{1}{\delta} \cdot \frac{C(1+\epsilon)}{D^*(\mu_1,\mu_2)}\right)\right].$$

By Lemma 8 below, we have that

$$\frac{nD^*(\mu_1,\mu_2)}{1+\epsilon} \geq \log\left(\frac{Cn}{\delta}\right) \; \forall\, n \geq N^*.$$

Therefore, under the event $\mathcal{E}_{N^*}(\eta^*)$, we have

$$\Lambda(N^*) \geq \log\left(\frac{CN^*}{\delta}\right)$$
$$\implies N_{\delta,\bar{\mu}} \leq \inf\left\{ n \in \mathbb{N} \;\middle|\; \Lambda(n) > \log\left(\frac{Cn}{\delta}\right)\right\} \leq N^*$$

where recall that $N_{\delta,\bar{\mu}}$ denotes the number of rounds at termination. In fact for any $n \geq N^*$, under the event $\mathcal{E}_n(\eta^*)$, we have that, $N_{\delta,\bar{\mu}} \leq n$. This means that $\mathbb{P}(N_{\delta,\bar{\mu}} > n) \leq \mathbb{P}(\mathcal{E}_n^C(\eta^*)) \,\forall\, n \geq N^*$. We will now bound this probability using Chernoff's bound ([30, Theorem 2.1 and Corollary 4.1]). For $n \geq N^*$,

$$\begin{aligned}\mathbb{P}(N_{\delta,\bar{\mu}} > n) &\leq \mathbb{P}(\mathcal{E}_n^C(\eta^*)) \\ &\leq \sum_{a=1}^{K} \mathbb{P}(|\hat{\mu}_a(n) - \mu_a| > \eta^*) \\ &\leq \sum_{a=1}^{K} \left[e^{-nd(\mu_a+\eta^*,\mu_a)} + e^{-nd(\mu_a-\eta^*,\mu_a)}\right].\end{aligned}$$

Now, we know that,

$$\begin{aligned}\mathbb{E}[N_{\delta,\bar{\mu}}] &= \sum_{n=0}^{\infty} \mathbb{P}(N_{\delta,\bar{\mu}} > n) \\ &\leq N^* + \sum_{n=N^*}^{\infty} \mathbb{P}(N_{\delta,\bar{\mu}} > n) \\ &\leq N^* + \sum_{n=N^*}^{\infty} \sum_{a=1}^{K} \left[e^{-nd(\mu_a+\eta^*,\mu_a)} + e^{-nd(\mu_a-\eta^*,\mu_a)}\right] \\ &\leq N^* + \sum_{a=1}^{K} \sum_{n=0}^{\infty} \left[e^{-nd(\mu_a+\eta^*,\mu_a)} + e^{-nd(\mu_a-\eta^*,\mu_a)}\right] \\ &\leq N^* + \sum_{a=1}^{K} \left[\frac{1}{1-e^{-d(\mu_a+\eta^*,\mu_a)}} + \frac{1}{1-e^{-d(\mu_a-\eta^*,\mu_a)}}\right].\end{aligned}$$

Since the right term only depends on ϵ and $\bar{\mu}$, we have that,

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1+\epsilon}{D^*(\mu_1,\mu_2)}.$$

Since, this is true for all $\epsilon > 0$, we have that:

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1}{D^*(\mu_1,\mu_2)}.$$

■

We now state the lemma we used in the proof above.

Lemma 8: For any two constants $c_1, c_2 > 0$,

$$x_0 = \frac{1}{c_1}\left[\log\left(\frac{c_2 e}{c_1}\right) + \log\log\left(\frac{c_2}{c_1}\right)\right]$$

is such that $c_1 x \geq \log(c_2 x) \,\forall\, x \geq x_0$.

Proof: Follows directly from [17, Lemma 18], combined with the fact that $c_1 x - \log(c_2 x)$ is increasing for $x > 1/c_1$. ■

Appendix F
Asmyptotic Optimality of PPR-JD

In this section, we first show that the PPR-JD stopping rule can also be expressed in terms of the Generalized Likelihood Ratio statistic $\Lambda(n)$ crossing a threshold. This will then allow us to apply Theorem 5 to prove the asymptotic optimality of PPR-JD.

We start with the form of PPR-JD stopping rule for $K = 2$ arms that we had derived in (4). Using Stirling's approximation [31], we have for $x \geq 1$,

$$\ln(x!) = x\ln(x) - x + \frac{1}{2}\ln(2\pi x) + O\left(\frac{1}{x}\right).$$

Substituting this for the factorial terms in our expressions and simiplifying we get,

$$\begin{aligned}&\ln(s_1^n!) + \ln(s_2^n!) + \ln(f_1^n!) + \ln(f_2^n!) - 2\ln((n+1)!) \\ &- (s_1^n + s_2^n)\ln(s_1^n + s_2^n) - (f_1^n + f_2^n)\ln(f_1^n + f_2^n) \\ &+ 2n\ln(2n) \\ &= n\left[d\left(\hat{\mu}_1, \hat{\mu}_{1,2}\right) + d\left(\hat{\mu}_2, \hat{\mu}_{1,2}\right)\right] + \ln(2\pi) \\ &+ \frac{1}{2}\ln(\hat{\mu}_1(1-\hat{\mu}_1)\hat{\mu}_2(1-\hat{\mu}_2)) + \ln(n) \qquad (9) \\ &- 2\ln((n+1)) + O\left(\frac{1}{n}\right). \qquad (10)\end{aligned}$$

Now, recall from (5) that

$$\Lambda_{1,2}(n) = n\left[d\left(\hat{\mu}_1, \hat{\mu}_{1,2}\right) + d\left(\hat{\mu}_2, \hat{\mu}_{1,2}\right)\right].$$

Also, define:

$$h_{1,2}(n) := \sqrt{\hat{\mu}_1(1-\hat{\mu}_1)\hat{\mu}_2(1-\hat{\mu}_2)}.$$

Substituting in (10) and using (4), we have that the PPR-JD stopping rule for two arms can be expressed as:

$$\begin{aligned}&\Lambda_{1,2}(n) + \ln(2\pi n) + \ln(h_{1,2}(n)) \\ &- 2\ln(n+1) + O\left(\frac{1}{n}\right) \geq \ln\left(\frac{1}{\delta}\right) \\ &\iff \Lambda_{1,2}(n) \geq \ln\left(\frac{n+1/n+2}{h_{1,2}(n)} \cdot \frac{e^{O(1/n)}}{2\pi\delta}\right). \qquad (11)\end{aligned}$$

Using Stirling's approximation [31], one can verify that $O(1/n)$ term $< 1/6n$ above. In addition, noting from before that we only need to check the PPR-JD rule for the top two empirical mean arms with error probability $\delta/(K-1)$, we

have the following lemma which relates the PPR-JD and the Chernoff stopping rules.

Lemma 9: For uniform sampling with K Bernoulli arms, the Chernoff stopping rule with threshold:

$$\beta(t,\delta) = \log\left(\frac{n+\frac{1}{n}+2}{h_{\alpha_1,\alpha_2}(n)} \cdot \frac{e^{1/6}}{2\pi} \cdot \frac{K-1}{\delta}\right) \tag{12}$$

is sufficient for the PPR-JD rule to be applied. That is, the PPR-JD algorithm stops before the Chernoff stopping rule with the above threshold is triggered.

Remark 10: Comparing the thresholds for the U-CNF and PPR-JD rules, we can see that the former has a K^2 dependence while the latter only has a linear dependence on K. This means we expect PPR-JD to perform better, which we do observe in our empirical results in Section V. However, one thing to note is that the $h_{\alpha_1,\alpha_2}(t)$ term in the threshold for PPR-JD becomes small when the mean rewards are near the extremes. This would suggest that the performance of PPR-JD will suffer in such cases, which also we are able to verify in our simulation results.

The above result also raises a more fundamental question of whether there is a connection between the PPR martingale methods and the Generalized Likelihood Ratio statistic methods. We were able to find a basic similarity between the two rules for Bernoulli Bandits, and we conjecture that there is possibly a similar connection for exponential family bandits as well.

Next, we provide the proof of the asymptotic optimality of PPR-JD in Corollary 6 which uses our derived sufficient threshold for PPR-JD in conjunction with the proof of Theorem 5. Due to the nature of $h_{\alpha_1,\alpha_2}(n)$ term, there are a few more details to take care of, which we furnish next.

Proof of Corollary 6: From Lemma 9 we have that the PPR-JD rule can be expressed in the form of $\Lambda(n)$ crossing a suitable threshold. Thus, we have

$$\begin{aligned} N_{\delta,\bar{\mu}} &\leq \inf\left\{n \in \mathbb{N} \;\middle|\; \Lambda(n) > \log\left(\frac{n+\frac{1}{n}+2}{h_{\alpha_1,\alpha_2}(n)} \cdot \frac{e^{1/6}}{2\pi} \cdot \frac{K-1}{\delta}\right)\right\} \\ &\leq \inf\left\{n > 2 \;\middle|\; \Lambda(n) > \log\left(\frac{2n}{h_{\alpha_1,\alpha_2}(n)} \cdot \frac{e^{1/6}}{2\pi} \cdot \frac{K-1}{\delta}\right)\right\}. \end{aligned}$$

We then proceed along the same lines as the proof of Theorem 5. We just need to additionally pick an η''' small enough such that,

$$h_{1,2}(n) = \sqrt{\hat{\mu}_1\hat{\mu}_2(1-\hat{\mu}_1)(1-\hat{\mu}_2)} \geq \frac{\sqrt{\mu_1\mu_2(1-\mu_1)(1-\mu_2)}}{2}$$

and we let $\eta_1^* = \min\{\eta', \eta'', \eta'''\}$. Then, we can define,

$$C' \coloneqq \frac{2e^{1/6}(K-1)}{\pi\sqrt{\mu_1\mu_2(1-\mu_1)(1-\mu_2)}}$$

and if we take,

$$\begin{aligned} N_1^* \coloneqq \frac{1+\epsilon}{D^*(\mu_1,\mu_2)} \Bigg[&\log\left(\frac{1}{\delta} \cdot \frac{C'e(1+\epsilon)}{D^*(\mu_1,\mu_2)}\right) \\ &+ \log\log\left(\frac{1}{\delta} \cdot \frac{C'(1+\epsilon)}{D^*(\mu_1,\mu_2)}\right)\Bigg] \end{aligned}$$

then, for any $N \geq N_1^*$ under the event $\mathcal{E}_N(\eta_1^*)$,

$$\begin{aligned} N_{\delta,\bar{\mu}} &\leq \inf\left\{n > 2 \;\middle|\; \Lambda(n) > \log\left(\frac{2n}{h_{1,2}(n)} \cdot \frac{e^{1/6}}{2\pi} \cdot \frac{K-1}{\delta}\right)\right\} \\ &\leq \inf\left\{n > 2 \;\middle|\; \Lambda(n) > \log\left(\frac{C'n}{\delta}\right)\right\} \\ &\leq N_1^*. \end{aligned}$$

As before, we have that

$$\begin{aligned} \mathbb{E}[N_{\delta,\bar{\mu}}] &\leq N_1^* + \sum_{n=N_1^*}^{\infty} \mathbb{P}(N_{\delta,\bar{\mu}} > n) \\ &\leq N_1^* + \sum_{a=1}^{K}\left[\frac{1}{1-e^{-d(\mu_a+\eta_1^*,\mu_a)}} + \frac{1}{1-e^{-d(\mu_a-\eta_1^*,\mu_a)}}\right]. \end{aligned}$$

Since the right term still only depends on ϵ and $\bar{\mu}$, we have that,

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1+\epsilon}{D^*(\mu_1,\mu_2)}.$$

Since, this is true for all $\epsilon > 0$, we have that:

$$\lim_{\delta\to 0} \frac{\mathbb{E}[N_{\delta,\bar{\mu}}]}{\log(1/\delta)} \leq \frac{1}{D^*(\mu_1,\mu_2)}.$$

■