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

Sriram Pillutla
ME21B196
Dept. of Mechanical Engineering, IIT Madras

Ruthwik Chivukula
ME21B166
Dept. of Mechanical Engineering, IIT Madras

***Abstract*—We consider the problem of best arm identification in stochastic multi-armed bandits, in the uniform sampling regime, which 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.**

***Index Terms*—Bernoulli bandits, pure exploration, stopping rule, PPR martingale, GLR, asymptotic optimality**

I. Introduction

In contrast with the classical task of regret-minimisation, Best-Arm Identification (BAI) is a problem of "pure exploration". BAI has been studied in two main settings. In the **fixed budget** setting, the algorithm has a given budget of T pulls for experimentation, to minimise either the probability of mis-identifying the best arm, or a related quantity called the "simple regret". 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. Our investigation is in the fixed confidence setting.

A. Contribution

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 propose two separate stopping rules for Bernoulli bandits. The first rule, denoted PPR-JD, is based on PPR martingale confidence sequences, which were recently also applied to the closely-related problem of PAC mode estimation. The second rule, denoted U-CNF, is based on the Chernoff rule. After presenting these stopping rules in Section V, we show that indeed both are asymptotically optimal in their sample complexity (while also being computationally efficient).

II. Related Work

A common aspect of several algorithms in the fixed-confidence setting is for sampling and stopping both to be guided by lower and upper confidence bounds on the unknown means of the arms. With upper bounds on sample complexity dependent on problem hardness, they are typically within a constant factor of the δ-dependent lower bound.

Garivier and Kaufmann 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 "TopTwo Thompson Sampling", "Top-Two Transportation Cost", and "BayesElim".

III. Problem Setup

We consider a stochastic bandit with $K \geq 2$ arms, denoted by $[K] := \{1, 2, \dots, K\}$. Each arm $a \in [K]$ has an associated distribution Π_a over scalar rewards, unknown to the learner. Random rewards $r \sim \Pi_a$ from the same arm a are i.i.d. samples from Π_a, with $\mathbb{E}[r] = \mu_a$.

To ensure a well-defined BAI problem, assume $\exists$ unique $a^* \in [K]$. Without loss of generality,

$$\mu_1 > \mu_2 \geq \mu_3 \geq \mu_4 \geq \cdots \geq \mu_K$$

Since we are considering the uniform sampling regime, 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.

For $\delta \in (0, 1]$, algorithm A is said to be δ-PAC if on every bandit instance ν, $P_\nu(\hat{a} \neq a^*) \leq \delta$. We seek δ-PAC algorithms that minimize $\mathbb{E}[N_{\delta,\nu}]$, which is the round complexity and K times the sample complexity.

A. Notation

A bandit instance ν fixes the probability distribution Π_a of each arm $a \in [K]$. Denote by Ω the set of all bandit instances.

Denote by $\mathcal{B}$ the set of all bandit instances where each arm's reward distribution is Bernoulli. We denote such a bandit instance by $\bar{\mu} := (\mu_1, \mu_2, \dots, \mu_K)$, which defines the instance $\Pi_a = \text{Bernoulli}(\mu_a)$, for all $a \in [K]$.

We also specifically define the term:

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

along with the additional convention that $d(0,0) = d(1,1) = 0$. Note that $d(\mu_a, \mu_b)$ 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, $\tau = NK$.

$\hat{\mu}_a$ is the empirical mean of arm a up to the current time τ or round n (implicit from context). We maintain $\hat{\mu}_{\alpha_1} \geq \hat{\mu}_{\alpha_2} \geq \cdots \geq \hat{\mu}_{\alpha_K}$, where α_i denotes the arm with the i-th highest empirical mean.

IV. Lower bounds on the round complexity

In this section, we present lower bounds on the round complexity of δ-PAC algorithms under uniform sampling.

Theorem 1 (General Lower Bound) *Let $\nu \in \Omega$ be any identifiable bandit instance with a unique best arm. Any δ-PAC uniform sampling algorithm on ν with stopping round $N_{\delta,\nu}$ satisfies*

$$E[N_{\delta,I}] \geq \frac{\log(1/2.4\delta)}{\inf_{I' \in \Omega'(I)} \sum_{a=1}^{K} KL(\Pi_a \| \Pi'_a)},$$

$\Omega'(\nu) = \{\nu' \mid a^*(\nu') \neq a^*(\nu),\ \nu' \in \Omega\}$ be the set of all identifiable bandit instances having a different best arm.

Corollary (Lower Bound for Bernoulli Bandits): Any δ-PAC uniform sampling algorithm with stopping round $N_{\delta,\bar{\mu}}$ satisfies

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

V. Algorithms

We propose these rules and analyse them in the context of Bernoulli bandits.

A. Prior-Posterior Ratio Martingale Based Stopping Rule

We take some initial prior over $\mathcal{P}$, say $f_0(p)$, with the goal of estimating $p^* \in \mathcal{P}$, the "ground truth" parameter. The posterior after t samples $X \equiv (X_1, X_2, \ldots, X_t)$ is given by

$$f_t(p) = \frac{f_0(p) L_p(X)}{\int_{q \in P} f_0(q) L_q(X)\, dq},$$

where $L_p(X)$ gives the likelihood of the outcomes for a given parameter p. Then, the prior-posterior ratio (PPR) at time t is the quantity

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

Waudby-Smith and Ramdas define

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

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

Proposition 3 (PPR Martingale): For any prior $f_0(p)$ on $\mathcal{P}$ that assigns non-zero mass everywhere, $(R_t(p))_{t=0}^{\infty}$ evaluated at p^* is a non-negative martingale with respect to $(F_t = \sigma(X))_{t=0}^{\infty}$ (in a way, the history of outcomes). Further,

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

Now, we work out a rule for BAI with uniform sampling.

1) $K = 2$ Bernoulli Arms: We will try to jointly estimate the two parameters (μ_1, μ_2). Since we need a prior with non-zero mass everywhere, a uniform prior is a suitable choice: $f_0(p_1, p_2) = 1$ for $p_1, p_2 \in [0,1]$.

After n rounds, suppose arm i has yielded s_n^i 1's and f_n^i 0's. Since the reward distributions are independent and Bernoulli, and have Beta distributions as conjugate priors,

$$f_n(p_1,p_2) = \text{Beta}(p_1; s_n^1 + 1, f_n^1 + 1)\,\text{Beta}(p_2; s_n^2 + 1, f_n^2 + 1)$$

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

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

To determine which arm dominates the other, it suffices to stop when C_n only contains points (p_1, p_2) such that either $p_1 > p_2$ or $p_2 > p_1$. Assume without loss of generality that arm 1 is empirically superior, i.e., $s_n^1 > s_n^2$. We look at the case where C_n contains "bad" points (p_1, p_2) such that $p_2 \geq p_1$.

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

Thus, if C_n contains "bad" points, it must contain at least one point of the form (p,p). Consequently, to stop, it suffices for C_n to separate from the line $p_1 = p_2$.

For $p \in [0,1]$, $f_n(p,p)$ can be represented as a single Beta pdf, with a multiplicative factor N_c independent of p:

$$f_n(p,p) = N_c \text{Beta}(p; s_n^1 + s_n^2 + 1, f_n^1 + f_n^2 + 1).$$

The mode of this Beta pdf occurs at $\hat{\mu}_{1,2} := \frac{s_n^1 + s_n^2}{2n} = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$.

Hence, our stopping rule can effectively be stated as: stop as soon as $f_n(\hat{\mu}_{1,2}, \hat{\mu}_{1,2}) \leq \delta$. Since we used the joint distribution over two arms for generating our confidence sequence, we call this rule PPR-JD.

2) $K \geq 2$ Bernoulli Arms: The PPR-JD stopping rule can be easily extended to instances with $K \geq 2$ arms by the "1 versus 1" approach. Let us assume without loss of generality that α_1 is the empirically best arm. To choose α_1 at any round n, $C_n^i = \left\{(p_{\alpha_1}, p_i) \in [0,1]^2 \mid f_n(p_{\alpha_1}, p_i) > \delta_K\right\}$ should contain only "good" points: $(p_{\alpha_1}, p_i), \quad p_{\alpha_1} > p_i \quad \forall i \in [K] \setminus \{\alpha_1\}$. It suffices to check the PPR-JD rule only with the $K-1$ pairs with $\delta_K = \delta/(K-1)$ to ensure that the overall confidence exceeds $1-\delta$.

Interestingly, it can be shown 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 condition simplifies to,

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

B. Chernoff's Stopping Rule for Uniform Sampling

The key is the Generalized Likelihood Ratio (GLR) statistic,

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

Intuitively, a higher $\Lambda_{a,b}(t)$ value places a higher belief on arm a being better than arm b. By definition, it follows that $\Lambda_{a,b}(n) = -\Lambda_{b,a}(n)$.

After adapting the analytical form of this term to uniform sampling, we get $\Lambda_{a,b}(n) = nD^*\left(\hat{\mu}_a(n), \hat{\mu}_b(n)\right)$ if $\hat{\mu}_a \geq \hat{\mu}_b$, where $D^*(\hat{\mu}_a, \hat{\mu}_b) := d\left(\hat{\mu}_a, \hat{\mu}_{a,b}\right) + d\left(\hat{\mu}_b, \hat{\mu}_{a,b}\right)$.

The stopping rule has an intuitive basis on this term with a specific choice of arms exceeding a threshold. We look for the arm a^* which maximizes $\Lambda_{a^*,b}$ (the best candidate for being the winner), and choose the other arm b^* such that it minimizes Λ_{a^*,b^*} (the toughest challenger to a^*). As this chosen term now crosses a threshold, it can be shown that the best arm can be identified with at least a $(1-\delta)$ probability.

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

Since $\Lambda_{a,b}(n)$ is only non-negative when arm a is empirically superior to arm b, the outer maximizer is clearly α_1 (if a was anything apart from α_1, the choice of b would be α_1, making the term negative). Combining this with the fact that $D^*(x, y)$ is decreasing in y for $x > y$, we obtain:

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

Garivier and Kaufmann 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:

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

Observe that both stopping rules only depend on the top two empirically superior arms.

C. Asymptotic Optimality of the Stopping Rules

Although PPR-JD and U-CNF are different rules, it is on account of satisfying the conditions of the following theorem that they become asymptotically optimal.

Theorem 2 (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*

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

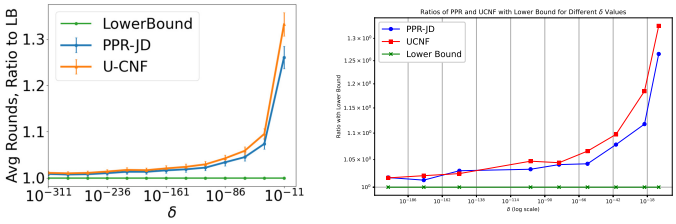
for some positive constant C, satisfies the following upper bound on its expected stopping time:

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

The upper bounds for Bernoulli bandits asymptotically match the lower bound from the corollary of Theorem 1.

VI. Main Results

The empirical results confirm that both the rules are asymptotically optimal with respect to the lower bound for Bernoulli bandits.



(a) From the paper (b) Our simulation

Fig. 1: Average rounds for PPR-JD and U-CNF with $\mu_1 = 0.7, \mu_2 = 0.5$

It was observed that for cases where the arms' means are distant from 0 and 1, PPR-JD terminates before U-CNF. This is apparent in the results from the original paper and our simulation. However, when the means are close to 0 and 1, U-CNF enjoys an advantage, albeit only moderately. This can be understood from the following term present in the denominator of the log term in the "threshold" for PPR-JD:

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

which is clearly small as the means approach the extremes. Therefore for most realistic bandit problems, PPR-JD appears to be the algorithm of choice.

VII. Conclusion

We examined the problem of Best Arm Identification (BAI) in a fixed confidence setting, utilizing a uniform sampling regime. We adapt Kaufmann et al.'s lower bound for uniform sampling and propose two stopping rules, PPR-JD and U-CNF, which are simpler and more efficient than fully sequential methods. Our analysis shows both rules are asymptotically optimal for Bernoulli bandits. The analytical results are reaffirmed by experiments, which also provide guidance for choosing between PPR-JD and U-CNF in practice. However, a gap remains between upper and lower bounds for finite δ values. Future work could explore a supplementary lower bound depending on K and extend these methods to other reward distributions.

VIII. Acknowledgement

We would like to express our sincere gratitude to Dr. Srinivas Reddy Kota for his invaluable guidance.