Optimal Stopping Rules for Best Arm Identification under Uniform Sampling

Sriram Pillutla, Ruthwik Chivukula

Indian Institute of Technology
Madras

28th October 2024

Introduction

- Best Arm Identification (BAI) has been studied in two main settings: fixed budget and fixed confidence.
- We focus on the fixed-confidence (fixed δ) setting with uniform sampling where the aim is to stop and correctly identify the best arm with probability at least $1 - \delta$.
- We will be covering two separate stopping rules for Bernoulli bandits specifically, namely, Prior Posterior Ratio (PPR-JD) and Uniform Chernoff (U-CNF).

Problem Setup

- A stochastic bandit with $K \geq 2$ arms, denoted $[K] := \{1, 2, \ldots, K\}$.
- Rewards $r \sim \Pi_a$ are i.i.d. with mean μ_a.
- Objective: Identify the best arm a^* with probability $1 - \delta$.

Lower Bound on Round Complexity

Theorem 1 (General Lower Bound): Any δ-PAC algorithm satisfies:

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

Corollary (Bernoulli Bandits):

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

Partial proof:

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

This can be simplified using the inequality $d(\delta, 1-\delta) \geq \log\left(\frac{1}{2.4\delta}\right)$.

PPR-JD Stopping Rule

- Prior-posterior ratio (PPR)

$$R_t(p) := \frac{f_0(p)}{f_t(p)}, \text{ where } f_t(p) = \frac{f_0(p)L_p(X)}{\int_{q\in\mathcal{P}} f_0(q)L_q(X)dq}.$$

- Waudby-Smith and Ramdas define

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

- This is a confidence sequence, because $(R_t(p))_{t=0}^{\infty}$ at $p = p^*$ forms a non-negative martingale w.r.t outcome history, and from Ville's inequality on martingales,

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

PPR: K=2 Bernoulli Arms

- We will try to jointly estimate (μ_1, μ_2), with a uniform prior.

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

- Confidence sequence for Bernoulli arms with two parameters, (μ_1, μ_2):

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

- Stop when C_n excludes points where $p_1 \leq p_2$ (or otherwise).

PPR: K=2 Bernoulli Arms

Lemma

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

Partial proof: Expanding $f_n(p_1, p_2)$ and using $\Delta := \bar{p} - p_1 = p_2 - \bar{p}$ to replace p_1 and p_2 gives the result.

- Thus, if C_n contains "bad" points, it must contain at least one point of the form (p, p).
- Set stopping to be when $f_n(\hat{\mu}_{1,2}, \hat{\mu}_{1,2}) \leq \delta$, as $\hat{\mu}_{1,2} := \frac{s_n^1+s_n^2}{2n} = \frac{\mu_1+\mu_2}{2}$ is the mode of $f_n(p, p)$.

Extension to K Arms

- PPR-JD rule extended to $K \geq 2$ arms using "1 versus 1" comparisons.
- Since the only possible winner is α_1, check only the $K-1$ pairs with $\delta_K = \delta/(K-1)$ to ensure the same $(1-\delta)$ confidence.
- Simplified rule: stop if $f_n(\hat{\mu}_{\alpha_1,\alpha_2}, \hat{\mu}_{\alpha_1,\alpha_2}) \leq \frac{\delta}{K-1}$. This comes from the result

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

U-CNF Stopping Rule

- The Generalized Likelihood Ratio (GLR) statistic is given as:

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

- Gives the log ratio of the maximum likelihood of **arm a** having a higher mean than **arm b** over the opposite hypothesis.
- This term can be further simplified under Bernoulli bandits and uniform sampling conditions as follows:

$$\Lambda_{a,b}(n) = nD^*(\hat{\mu}_a(n), \hat{\mu}_b(n))$$

where

$$D^* := d\left(\hat{\mu}_a(n), \frac{\hat{\mu}_a(n) + \hat{\mu}_b(n)}{2}\right) + d\left(\hat{\mu}_b(n), \frac{\hat{\mu}_a(n) + \hat{\mu}_b(n)}{2}\right)$$

U-CNF Stopping Rule

- Propose stopping iff

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

- Stop and return α_1 if:

$$nD^*(\hat{\mu}_{\alpha_1}(n), \hat{\mu}_{\alpha_2}(n)) > \beta(n, \delta)$$

- Garivier and Kaufmann provide an informational threshold $\beta(n, \delta)$ for Bernoulli bandits, which they prove to be δ-PAC for any sampling strategy. Choose

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

Asymptotic Optimality

Theorem (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:

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

Both PPR-JD and U-CNF are asymptotically optimal as $\delta \to 0$.

Asymptotic Optimality

- Evidently, the U-CNF rule satisfies the Asymptotic Upper Bound theorem by setting $C = 2K(K-1)$.
- For PPR-JD, it is not very obvious. But it can be shown by invoking the Stirling approximation.

$$\Lambda_{1,2}(n) \geq \ln\left(\frac{\frac{n+1}{n+2}}{h_{1,2}(n)} \cdot \frac{e^{O(1/n)}}{2\pi\delta}\right)$$

- where $h_{1,2}(n)$ is given as:

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

Empirical Results

Results with both rules for $\mu_1 = 0.7$ and $\mu_2 = 0.5$.

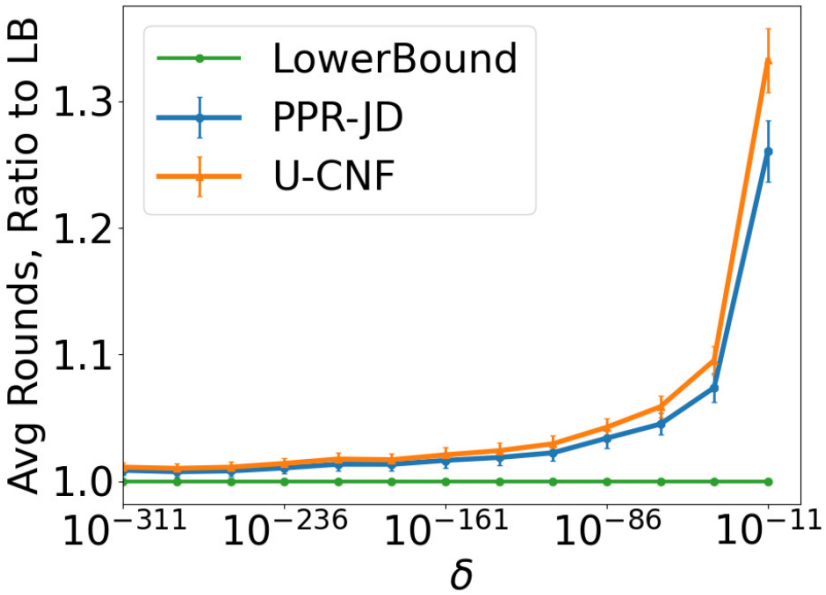

Figure 1: Results from the paper.

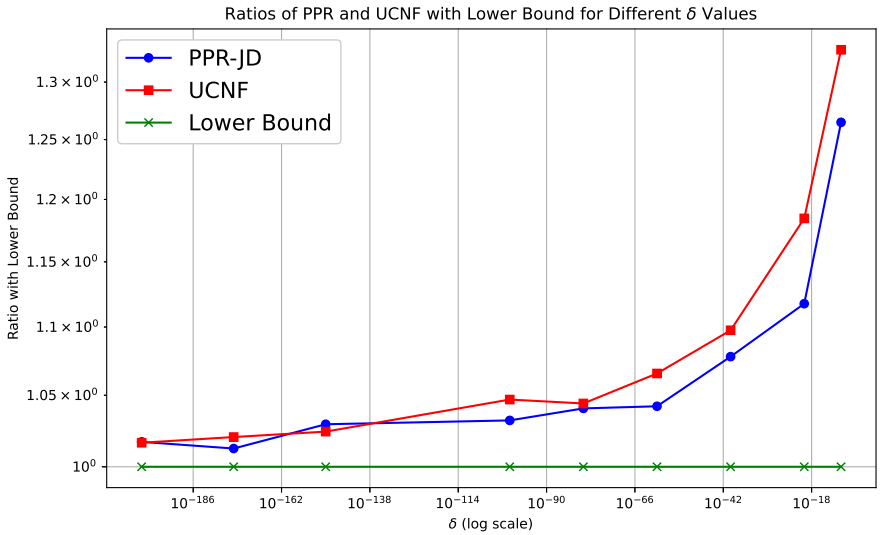

Figure 2: Our simulation.

Empirically we confirm the analytical results, showing asymptotic optimality for both. Notably, U-CNF performs better than PPR-JD when arms' means are at the extremes.

Conclusion

- Proposed two stopping rules for BAI in Bernoulli bandits: PPR-JD and U-CNF.
- Both rules are computationally efficient and asymptotically optimal.
- Choice depends on properties of bandit instance (means close vs. extremes).