Best-arm Identification in Best of Both Worlds Setting

Adhyyan Narang Yunkyu Song Claire (Jie) Zhang

June 2021

1 Introduction

In the multi-armed bandit problem, at each step, the learning algorithm chooses one arm (one action), and receives a reward. The most commonly studied problem is that of regret minimization i.e maximizing rewards with respect to some fixed baseline. However, the focus of this paper is on the problem of best-arm identification. This is different from regret minimization problem in the sense that the algorithm's performance will be evaluated by its ability to determine the best-arm. If the algorithm chooses the second-best arm or the worst arm, they still get the same error by this metric, unlike regret.

The paper [1] focuses on the variant of best-arm identification problem with fixed-budget of n pulls. Later in the related works section of the scribe, a couple of other variants of the same problem will be mentioned . Previous works ([2] and a few others)on the same setting and variant had focused on stochastic rewards, [1] is the first to consider the best-arm identification problem in adversarial setting and thus the best of both worlds problem.

In the stochastic setting, best-arm is the arm with maximum mean rewards.

$$k^* = \arg\max_{k \in [K]} \mu_k \tag{1}$$

In the adversarial setting, suppose $\boldsymbol{g}$ is a reward matrix for all k arms and n rounds pulled, best arm is defined to be the arm with maximum cumulative rewards:

$$k^* = \arg\max_{k \in [K]} \sum_{t=1}^{n} \boldsymbol{g}_{k,t} \tag{2}$$

Note that interaction protocal in both cases is the same as in other multi-armed bandit problems: given round t, each arm k is assigned reward $\boldsymbol{g}_{k,t}$, learner chooses to see $\boldsymbol{g}_{I_t,t}$.

Note that the adversarial case of best-arm identification has different motivations as the stochastic case. In the stochastic rewards case, it is assumed that each arm has a fixed distribution. Thus identifying the best arm means keep pulling that same arm after the n allocated exploration rounds. And rewards are assumed to be maximized if the arm is identified correctly. In the adversarial case, there is no assumption of future rewards similar to past rewards. Motivation in this case assumed action to be taken according to the information, i.e. gathered during the n pulls. One real life application given in [1] is in law enforcement data collection and decision making. Suppose law enforcement agencies have to monitor several criminal targets during a year, and at each end of year to decide on which target to act on. Suppose the law enforcement agencies have limited resources and only observe the activities of one criminal target they closely monitor at one time. During the year, they want to be robust to adversarial criminals (which maybe during the earlier month of the year do not carry out much activities to obscure the decisions of law enforcement). Real life application for stochastic case is discussed in [2].

1.1 Related Works

Best arm identification problem in the stochastic case has been studied in a few different variants. The fixed-budget set ting assumes the available pulls is given as n, and at end of n rounds, learning algorithm output best arm J_n. The probability of outputting a wrong arm is a measure of success. Prior to paper [1], there are a series of papers. [4] noted that in the pure exploration phase, there is still a trade-off between exploration and exploitation. One of the main focus of this paper is to characterize relation between cumulative regret in the pure exploration steps n, with simple regret (see Section 2). It also discussed different allocation strategies (I_n), combined with different recommendation strategies (J_n) in the stochastic case and upper and lower bounds on their error probability (see Section 2). Following that, [2] introduced a strategy Successive Rejects which enjoys a nice upper bound. Follow-up works [7] and [5] had given both tighter upper bounds and lower bounds. We note that all of those previous papers on the best-arm identification problem with fixed budget pulls focus on stochastic case. [1] is the first to focus on best of both worlds (BOB) setting.

In the study of multi-armed bandit problems, minimizing cumulative regret is often a major goal. Best of both worlds setting have been studied in this setting by [3] and many follow up works. On the other hand, the non-stochastic variant of best-arm identification has been studied by [6] and [8]. The non-stochastic setting in those two papers are different from the adversarial setting in [1] in that the non-stochastic rewards are assumed to converge when time horizon goes to infinity. Another important variant of the best-arm identification problem is studied in the setting of fixed confidence. In this setting, learning algorithm is given target confidence level for its best arm output, and is expected to use as less pulls as possible. In [7], both fixed budget setting and fixed confidence upper bounds on stochastic setting are provided.

2 Problem Setup

Notations We denote the available allocated pulls (time horizon) as n. The arms are denoted by $[K]$. In the stochastic case, they have distributions $\nu_1, \ldots, \nu_K$, with mean $\mu_1, \ldots, \mu_K$. The distributions and means are unknown to learning algorithm. At the end of n rounds, learner is expected to identify an arm J_n that has minimum simple regret r_n,

$$r_n = \max_{k \in [K]} \mu_k - \mu_{J_n}$$

Similar to other multi-armed bandits problem, we define gap Δ_i for arm $i \neq i^*$ as $\Delta_i = \mu^* - \mu_i$. Gap for best arm i^* is defined as $\Delta_{i^*} = \min_{i \neq i^*} \Delta_i$. We assume the indices of the arms follow the order such that $\Delta_{i^*} = \Delta_{(1)} = \Delta_{(2)} \leq \Delta_{(3)} \leq \cdots \leq \Delta_{(K)}$.

In the analysis of [1] and [2], algorithms are evaluated by the probability of error $e_n = \mathbb{P}(J_n \neq k^*)$. We note that in the stochastic case, expected simple regret is upper bounded by e_n. Given

$$\mathbb{E}(r_n) = \sum_{i \neq i^*} \mathbb{P}(J_n = i)\Delta_i$$

we have

$$\Delta_{i^*} e_n \leq \mathbb{E}(r_n) \leq e_n.$$

Therefore, it makes sense to bound e_n.

In the adversarial case, best arm is the arm with maximum cumulative rewards. We define gap in the adversarial case as:

$$n\Delta_k^g \triangleq \begin{cases} G_{(1)} - G_k, & \text{if } k \neq k_g^* \\ G_{(1)} - G_{(2)}, & \text{if } k = k_g^* \end{cases} \tag{3}$$

Estimators We note that in both the stochastic case and adversarial case, an estimation for cumulative rewards for each arm was needed to decide both which arm to pull and which arm to recommend. In the stochastic case, empirical mean estimator is commonly used. It is usually written in this form:

$$\hat{G}_k \triangleq \frac{n \sum_{t=1}^{n} 1[I_t = k] g_{k,t}}{\sum_{t'=1}^{n} 1[I_{t'} = k]} \tag{4}$$

In the adversarial case, an unbiased but potentially high variance estimator is commonly used:

$$\widetilde{G}_k \triangleq \sum_{i=1}^{n} \frac{g_{k,t}}{p_{k,t}} 1_{[I_t=k]} \tag{5}$$

In the analysis and discussion algorithms to follow, we will frequently refer to those two estimators.

Notions of Complexity Previous papers on the best-arm identification problem have come up with metrics that relates the error probability to a few of the quantities that are problem specific, such as gaps $\Delta_1, \dots, \Delta_k$ and number of arms K. Those quantities characterize how difficult it is to identify the best arm. Intuitively, having many arms means need more pulls to explore them. And having a small gap implies it is hard to distinguish the arms. In the following sections, we will see that a metric defined as equation 6 is a good characterization for the adversarial case. For the stochastic case, it is shown in the paper [2] that equation 7 is a good metric.

$$H_{\text{UNIF}} = \frac{K}{\Delta_{(1)}^2} \tag{6}$$

$$H_{\text{SR}} = \max_k \frac{k}{\Delta_k^2} \tag{7}$$

To characterize the hardness of the best of both worlds problem, we will come up with new metric in the sections to follow.

Class of problems To ease the analysis and discussion given the rewards could be non-stochastic, authors of [1] grouped problems according to their gaps into problem classes. Classes are denoted as $\boldsymbol{\Delta}_c$ with c being an positive integer. We now give the formal definition.

Definition 2.1. *In adversarial case, g is said to be in problem class $\boldsymbol{\Delta}_c$, if for all $k \in [K]$ except one arm $\bar{k}$, $\Delta_k/c \leq \Delta_k^g \leq c\Delta_k$, and for $\bar{k}$, its gap is related to the smallest gap: $\Delta_1/c \leq \Delta_{\bar{k}}^g \leq c\Delta_1$.*

3 RULE algorithm

In this section, we present a naive algorithm that samples arms uniformly at random. Then we characterize its adversarial identification error, and show that it must be optimal in an order-sense. We highlight the key insights that comprise the proofs. We present this simpler case in-depth because understanding the behavior of the Rule algorithm deeply is helpful to motivate the design and analysis of the $P1$-Algorithm that the authors propose.

Algorithm. At each round:

1. Rule selects an arm $I_t \in [K]$ uniformly at random.
2. Gets some gain $g_{I_t,t}$.

Then, at the end of the game, Rule computes $\widetilde{G}_{k,n}$ for all arms: the importance-weighted cumulative gain estimator, and makes a guess for the best arm as

$$J_n = \arg\max_k \widetilde{G}_{k,n}$$

The two theorems below show that the adversarial misidentification error obtained by Rule is optimal in an order-sense.

Theorem 1 (Theorem 1 in original)**.** *Fix the time horizon n, and let rewards $\boldsymbol{g}$ be chosen by an oblivious adversary, $\boldsymbol{g}_{k,t} \in [0,1]$. Then, for all $t \in [n]$ and for all $k \in [K]$, RULe outputs an arm J_n with the guarantee that its probability of error $e_{\mathrm{ADV}(g)}(n)$ verifies*

$$e_{\mathrm{ADV}(g)}(n) \leq K \exp\left(-\frac{3n}{28 H_{\mathrm{UNIF}}(\boldsymbol{g})}\right)$$

Below, we break-down the proof of the theorem into more understandable pieces, and highlight its connection to classical arguments in information theory.

Proof. First, we assume WLOG that arm 1 is the best. Hence, the probability of error is the probability of choosing any arm other than the first.

$$e_{\mathrm{ADV}(g)}(n) \triangleq \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}}\right) = \mathbb{P}\left(\exists k \in [2:K] : \widetilde{G}_{1,n} \leq \widetilde{G}_{k,n} \mid \boldsymbol{g}\right)$$

There are two "bad events" that could cause an error:

1. Estimated rewards for k_{th} arm are very high: $\widetilde{G}_{k,n} \geq G_k + \frac{n\Delta^g_k}{2}$.
2. Estimated rewards for 1_{st} arm is very low: $\widetilde{G}_{1,n} \leq G_1 + \frac{n\Delta^g_1}{2}$

In the absence of these, misclassification is impossible. Recognize that isolating these bad events to bound these probabilities independently is a technique that is native to Information Theory via the proof of the Shannon-Coding theorem; this technique is often observed in the analysis of randomized algorithms. We use a union bound to split the expression above as:

$$\begin{aligned}
&\leq \mathbb{P}\left(\exists k \in [2:K] : \widetilde{G}_{k,n} - G_k \geq \frac{n\Delta^g_k}{2} \text{ or } \widetilde{G}_{1,n} - G_1 \leq \frac{n\Delta^g_1}{2} \mid \boldsymbol{g}\right) \\
&\leq \mathbb{P}\left(\widetilde{G}_{1,n} - G_1 \leq \frac{n\Delta^g_1}{2} \mid \boldsymbol{g}\right) + \sum_{k=2}^{K} \mathbb{P}\left(\widetilde{G}_{k,n} - G_k \geq \frac{n\Delta^g_k}{2} \mid \boldsymbol{g}\right)
\end{aligned}$$

To bound each of these probabilities, we need an appropriate concentration inequality. The Bernstein inequality below is ideal for our purposes.

Lemma 1.1 (Bernstein inequality)**.** *Let X_i be a sequence of centered independent random variables:*$\mathrm{E}\left(X_i\right) = 0, \quad |X_i| \leq R, \quad \sigma^2 = \sum_{i=1}^n \mathbb{E}X_i^2$*, we have:*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2/2}{Rt/3 + \sigma^2}\right)$$

To apply Bernstein to the present case, we write $\widetilde{G}_{k,t} = \sum_{t=1}^n \widetilde{g}_{k,t} - g_{k,t} = \sum_{t=1}^n d_{k,t}$. All $d_{k,t}$ are centered. Then, $\widetilde{g}_{k,t} = \frac{g_{k,t}\mathbf{1}(I_t=k)}{p_{k,t}}$ is Bernoulli with param $1/K$ and range $[0, Kg_{k,t}]$. For the variance, $\sigma^2_{d_{k,t}} = \sigma^2_{\widetilde{g}_{k,t}} = \frac{1}{K}(1-\frac{1}{K})g^2_{k,t}$.

Applying the Bernstein to each k in the summation, we obtain

$$\begin{aligned}
\mathbb{P}\left(\widetilde{G}_{k,n} - G_{k,n} \geq \frac{n\Delta^g_k}{2}\right) &\leq \exp\left(-\frac{\left(\Delta^g_k/2\right)^2 n^2/2}{\sum_{t=1}^n \sigma^2_{d_{k,t}} + \frac{1}{6}K\Delta^g_k n}\right) \\
&\leq \exp\left(-\frac{\left(\Delta^g_k\right)^2 n^2/8}{nK + \frac{1}{6}Kn}\right) \\
&= \exp\left(-\frac{3\left(\Delta^g_k\right)^2 n}{28K}\right)
\end{aligned}$$

Notice the K in denominator. This is what causes the H_{UNIF} to show up in the theorem. Substituting back into the error-bound completes the proof. □

Now the natural question is whether some more intelligent adaptive-sampling scheme could do better. The theorem below shows that this is not possible.

Theorem 2. *Consider any problem class Δ_3 with associated complexity H_{UNIF}. For any learner, for any horizon n such that $K \exp\left(-n\Delta_1^2/128\right) \leq 1/128$ and $K \geq 4096$, there exist $\boldsymbol{g}^1 \in \boldsymbol{\Delta}_3$ and $\boldsymbol{g}^2 \in \boldsymbol{\Delta}_3$ such that:*

$$\max\left(e_{\boldsymbol{g}^1}(n), e_{\boldsymbol{g}^2}(n)\right) \geq \min\left(\frac{1}{128}\exp\left(-\frac{32n}{H_{\text{UNIF}}}\right), \frac{1}{32}\right)$$

The theorem is interesting: to show the existence of a single game, we consider two games together and show that at the algorithm must achieve bad error on at least one of these. The spirit of the theorem is reminiscent of the Theorem of the Alternatives in optimization; however, the proof, sketched below, does not follow from dual optimization problems.

First, we recall the Pigeonhole Principle: "if n items are put into m containers, with $n > m$, then at least one container must contain more than one item." This innocuous-looking result from elementary probability ends up being tremendously useful for our purposes.

Proof Sketch. Consider a "base game" with means $\mu_1 = \frac{1}{2}$ and $\mu_k = \frac{1}{2} - \Delta_k$. Consider the first $n/2$ rounds of the game. By a variant of the pigeonhole principle, we have at least one arm, indexed by $\bar{k}$ that is pulled less than $n/(2K)$ times. Now, we construct two very similar games from the base. For both games, $\forall k \neq \bar{k}, g^1_{k,t} = g^2_{k,t}$.

1. Game 1: For first half, follow Bernoulli exactly. For second half, rewards of all arms are 0, except $\bar{k}$ which is 1. This makes it the second best arm with expected total reward $n(\frac{1}{2} - \Delta_1)$.
2. Game 2: Second half exactly the same as above. For the first half, $\mu_{\bar{k}} = \frac{1}{2} - \Delta_{\bar{k}} + 2\Delta_1$. This makes expected reward $n(\frac{1}{2} + \Delta_1)$, and hence the best arm.

For the learner, since $\bar{k}$ is undersampled in L, this difference is not detected with high probability! To show this, the following lemma is used, whose proof is contained in the original paper.

Lemma 2.1 (Lemma 10 in original)**.** *Let L be a phase, i.e., a subset of rounds of the game, $L \subset [n]$.*

- *Consider two bandit problems. In both, $\forall t \in [n], \forall k \in [K]$, the rewards $\boldsymbol{g}_{k,t} \sim \nu_{k,t}$i.i.d. The two problems only differ in rewards for arm $\bar{k}$ during phase L.*
- *$\bar{k}$ is distributed as Bernoulli with means $\mu^{2}_{\frac{1}{\bar{k}}}(t) \triangleq \mu^{2}_{\bar{k}} \triangleq 1/2 + \Delta$ and $\mu_{\frac{1}{\bar{k}}}(t) \triangleq \mu^{1}_{\bar{k}} \triangleq 1/2 - \Delta'$ respectively for the two problems, where $1/8 > \Delta' \geq \Delta \geq 0$.*
- *We have an event W depending only on g generated by the problems and the actions of the learner $I_{[n]}$*

When $\bar{k}$ is pulled during phase L less than B times, we have

$$\mathbb{P}_2(W) \geq \frac{\mathbb{P}_1(W)}{8}\exp\left(-16\left(\Delta'\right)^2 B\right)$$

□

4 The goal moving forward: Best of Both Worlds

The BOB criterion: Consider the criterion defined below to be a best-of-both-worlds estimator. It requires that in an order-sense the algorithm both does as well as the Successive-Rejects algorithm for the stochastic problem and as well as Rule for the adversarial problem.

$$e_{\text{STO}}(n) \leq \widetilde{\mathcal{O}}\left(\exp\left(-\frac{n}{H_{\text{SR}}\log K}\right)\right) \quad \text{and} \quad e_{\text{ADV}(g)}(n) \leq \widetilde{\mathcal{O}}\left(\exp\left(-\frac{n}{H_{\text{UNIF}}(\boldsymbol{g})}\right)\right) \tag{8}$$

The question the rest of the paper is concerned with is whether this criterion is attainable. To understand why it is a hard problem, recall our two estimators. Defining $\widetilde{g}_{k,t} = \frac{g_{k,t} 1_{[I_t=k]}}{p_{k,t}}$,

$$\hat{G}_k \triangleq \frac{n \sum_{t=1}^{n} 1_{[I_t=k]} g_{k,t}}{\sum_{t'=1}^{n} 1_{[I_{t'}=k]}} \text{ and } \widetilde{G}_{k,t} = \sum_{t'=1}^{n} \widetilde{g}_{k,t'} \tag{9}$$

The first is unbiased in the stochastic case, yields optimal bounds. But biased in the adversarial case. The second has high variance for stochastic. Hence, naive adoption of neither is able to fulfill BOB. If there exists an algorithm that solves the BOB problem, it must intelligently reduce the bias of the first or the variance of the second. Theorem 3 states that, in the form stated above, BoB is unattainable.

***H_{BOB}* and comparison with previous measures** But before we can understand the theorem, we must introduce a new complexity measure:

$$H_{\text{BOB}} \triangleq \frac{1}{\Delta_{(1)}} \max_{k\in[K]} \frac{k}{\Delta_{(k)}}$$

Let us first compare this new measure with H_{SR}.

$$H_{\text{BOB}} = \frac{1}{\Delta_{(1)}} \max_{k\in[K]} \frac{k}{\Delta_{(k)}} \geq \max_{k\in[K]} \frac{k}{\Delta_{(k)}^2} = H_{\text{SR}} \tag{10}$$

Equality holds when the 2^{nd} arm is the argmax. We can now compare with H_{UNIF}.

$$H_{\text{BOB}} = \frac{1}{\Delta_{(1)}} \max_{k\in[K]} \frac{k}{\Delta_{(k)}} \leq \frac{K}{\Delta_{(1)}} \max_{k\in[K]} \frac{1}{\Delta_{(k)}} \leq \frac{K}{\Delta_{(1)}^2} = H_{\text{UNIF}} \tag{11}$$

Equality holds when the $\Delta_1 = \Delta_K$ i.e all same Δ's. To interpret when these inequalities are strict, we consider different cases.

1. $H_{SR} = H_{BOB} = H_{UNIF}$: In this case, Rule achieves the Bob criterion because the stochastic and adversarial inequality have the same order. Figure 1a illustrates that this can only happen when all Δ's are exactly equal, an unlikely scenario.

2. $H_{SR} = H_{BOB} < H_{UNIF}$: In this case, BoB is achievable but not by Rule. See Figure 1b.

3. $H_{SR} < H_{BOB} < H_{UNIF}$: This is the most general case and it is easy to choose Δ's such that neither Inequality 10 or Inequality 11 above is tight, and $H_{BOB} = \sqrt{K/2} H_{SR}$.

Now, we are prepared to understand the statement of the theorem below. Because of Inequality 10, one consequence of the theorem is that the BOB criterion stated as in Equation 8 is unattainable by any algorithm.

Theorem 3. *For any class problem Δ_4, for any learner, $\exists$ an i.i.d. stochastic problem* STO $\in \Delta_4$ *with complexity H_{BOB}, such that for any n satisfying $K \exp\left(-\Delta_1^2 n/32\right) \leq 1/32$, if*

$$e_{\text{STO}}(n) \leq \frac{1}{64} \exp\left(-\frac{2048n}{H_{\text{BOB}}}\right)$$

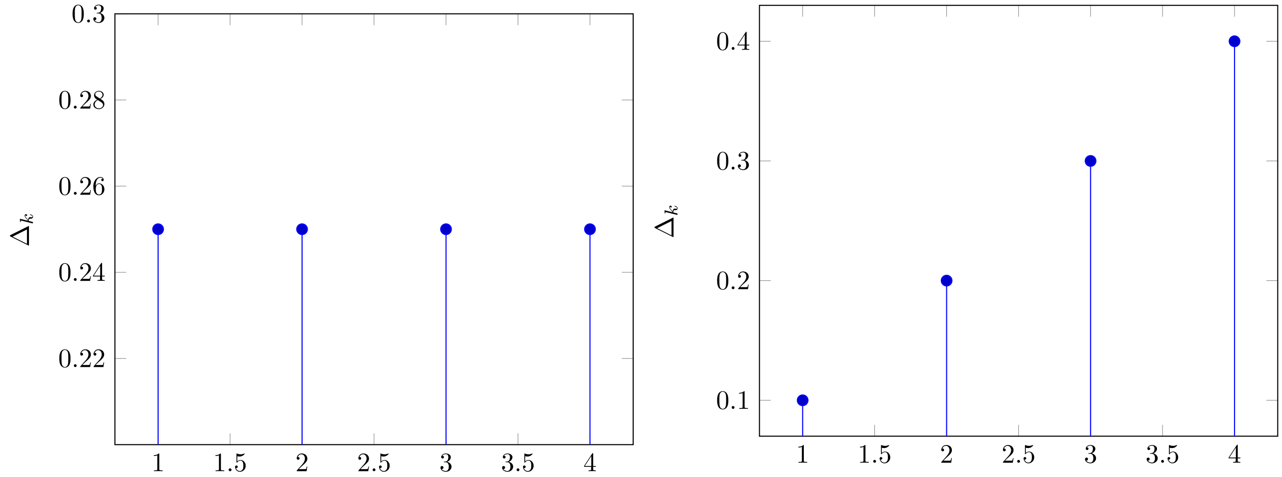
Then there exists an adversarial problem $\boldsymbol{g} \in \boldsymbol{\Delta}_4$ such that

$$e_{\text{ADV}(g)}(n) \geq \frac{1}{16}$$

The proof of the theorem is actually now being reviewed by an author.

5 P1 algorithm

In this section, we provide the main algorithm of the paper.



(a) Case 1

(b) Case 2

Figure 1: Illustration of different cases of equality between the three proposed complexity measures.

Algorithm. At each round:

1. Sorts arms in a decreasing order by $\widetilde{G}_{k,t}$, and denote the rank of the kth arm as $\widetilde{\langle k\rangle}_t$
2. P1 selects arm $I_t \in [K]$, with $p_{k,t} = P(I_t = k) = \frac{1}{\widetilde{\langle k\rangle}_t\overline{\log}(K)}$

As in RULE, the algorithm recommends

$$J_n = \arg\max_k \widetilde{G}_{k,n}$$

at the end of the game.

The performance of the algorithm is given in the theorem below.

Theorem 4. *In the stochastic case, the algorithm satisfies*

$$e_{STO}(n) \leq 2K^3 n e^{-\frac{n}{128 H_{P1}}}$$

and in the adversarial case, the algorithm satisfies

$$e_{ADV}(n) \leq K e^{-\frac{3n}{40\overline{\log}(k) H_{UNIF(\boldsymbol{g})}}}$$

where H_{P1} is a complexity in $O(H_{BOB}\log^2(K))$ defined as

$$H_{P1} \equiv \min_{\boldsymbol{a}\in\boldsymbol{A}} \max_{k\in K} \frac{\sum_{i=\langle k\rangle}^{K} (a_i - a_{i+1})i + \frac{1}{24}K a_{\langle k\rangle}\Delta_k}{a_{\langle k\rangle}^2 \Delta_k^2} \text{ where } \boldsymbol{A} \text{ is given as}$$

$$\boldsymbol{A} \equiv \{\boldsymbol{a} \in [0,1]^K : na_i \in \mathbb{N} \text{ for all } i \in [K], 1 = a_1 = a_1 \geq \ldots \geq a_K > 0\}$$

The algorithm is similar to algorithms that work well in each settings, in that it uses an IPS estimator and recommends an arm by computing the arg max as in RULE, which works well in the adversarial case, and it uses a weighting based on $\overline{\log}(K)$ as in SR, which works well in the stochastic case.

By drawing each arm with a probability of $\frac{1}{\widetilde{\langle k\rangle}_t\overline{\log}(K)}$, it works well on the stochastic case, because the lowest probability we draw the arm is bounded below by $\frac{1}{K\overline{\log}(K)}$ and this provides a bound to the variance of each

$\boldsymbol{g}_{k,t}$ term within the IPS estimator we are using, which further bounds the probability that the algorithm does not work well in each "step," where the step here is defined similarly in SR. On the other hand, the algorithm works well in adversarial case, since the probability we draw arms does not differ significantly from the uniform distribution, which we use in RULE algorithm.

Although we mainly discussed about the problem of identifying the best arm given a fixed budget under the setting that is either stochastic or non-adaptive adversarial, the P1 algorithm can be used in the follow settings and problems as well:

1. Fixed confidence interval: given a confidence value δ, a learner stops as soon as possible and returns the estimated best arm that is correct with a probability of at least $1 - \delta$.

2. Streams, windows, thresholds: a learner recommends the best arm in the latest time window between $t - W$ and t for each round t.

3. m-sets: a learner recommends m best arms.

4. Active anomaly detection: a learner monitors different streams of non-stochastic rewards and could potentially detect an anomaly if one of the streams outputs a reward signal that is on average larger than a given threshold during a time window period W.

5. Adaptive adversary (given a condition that H_{UNIF} is an upper bound on the complexity of all $\boldsymbol{g}$ that the adaptive adversary can possibly generate): an adversary chooses a reward for each arm depending on the algorithm and what arms have been chosen so far.

References

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic adversarial best-arm identification. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 918–949. PMLR, 06–09 Jul 2018.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

[3] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[4] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[5] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[6] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pages 240–248. PMLR, 2016.

[7] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

[8] Lisha Li, Kevin Jamieson, Giulia De Salvo, Rostamizadeh A Talwalkar, and A Hyperband. A novel bandit-based approach to hyperparameter optimization. *Computer Vision and Pattern Recognition, arXiv: 1603.0656*, 2016.