Best Arm Identification in Multi-Armed Bandits

Jean-Yves Audibert, Sébastien Bubeck

Best Arm Identification in Multi-Armed Bandits

Jean-Yves Audibert
Imagine, Université Paris Est
&
Willow, CNRS/ENS/INRIA, Paris, France
audibert@imagine.enpc.fr

Sébastien Bubeck, Rémi Munos
SequeL Project, INRIA Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France
{sebastien.bubeck, remi.munos}@inria.fr

Abstract

We consider the problem of finding the best arm in a stochastic multi-armed bandit game. The regret of a forecaster is here defined by the gap between the mean reward of the optimal arm and the mean reward of the ultimately chosen arm. We propose a highly exploring UCB policy and a new algorithm based on successive rejects. We show that these algorithms are essentially optimal since their regret decreases exponentially at a rate which is, up to a logarithmic factor, the best possible. However, while the UCB policy needs the tuning of a parameter depending on the unobservable hardness of the task, the successive rejects policy benefits from being parameter-free, and also independent of the scaling of the rewards. As a by-product of our analysis, we show that identifying the best arm (when it is unique) requires a number of samples of order (up to a $\log(K)$ factor) $\sum_i 1/\Delta_i^2$, where the sum is on the suboptimal arms and Δ_i represents the difference between the mean reward of the best arm and the one of arm i. This generalizes the well-known fact that one needs of order of $1/\Delta^2$ samples to differentiate the means of two distributions with gap Δ.

1 Introduction

In the multi-armed bandit problem, at each stage, an agent (or forecaster) chooses one action (or arm), and receives a reward from it. In its stochastic version, the reward is drawn from a fixed probability distribution given the arm. The usual goal is to maximize the cumulative sum of rewards, see Robbins (1952); Auer et al. (2002) among many others. Since the forecaster does not know the distributions, he needs to explore (try) the different actions and yet, exploit (concentrate its draws on) the seemingly most rewarding arms. In this paper, we adopt a different viewpoint. We assume that after a given number of pulls, the forecaster is asked to output a recommended arm. He is then *only* evaluated by the average payoff of his recommended arm. This is the so-called pure exploration problem, Bubeck et al. (2009).

The distinguishing feature from the classical multi-armed bandit problem described above is that the exploration phase and the evaluation phase are separated. Thus, there is no explicit trade-off between the exploration and the exploitation while pulling the arms. The target of Hoeffding and Bernstein races, see Maron and Moore (1993); Mnih et al. (2008) among others, is more similar to ours. However, instead of trying to extract from a fixed number of rounds the best action, racing algorithms try to identify the best action at a given confidence level while consuming the minimal number of pulls. They optimize the budget for a given confidence level, instead of optimizing the quality of the recommendation for a given budget size. Another variant of the best arm identification task is the problem of minimal sampling times required to identify an ε-optimal arm, see in particular Domingo et al. (2002) and Even-Dar et al. (2006).

We now illustrate why this is a natural framework for numerous applications. Historically, the first occurrence of multi-armed bandit problems was given by medical trials, see Robbins (1952). In the case of a severe disease, ill patients only are included in the trial and the cost of picking the wrong treatment is high. It is important to minimize the cumulative regret, since the test and cure phases coincide. However, for cosmetic products, there exists a test phase separated from the commercialization phase, and one aims at minimizing the regret of the commercialized product rather than the cumulative regret in the test phase, which is irrelevant.

Another motivating example concerns channel allocation for mobile phone communications. During a very short time before the communication starts, a cellphone can explore the set of channels to find the best one to operate. Each evaluation of a channel is noisy and there is a limited number of evaluations

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions $\nu_1, \dots, \nu_K$ of the arms.

For each round $t = 1, 2, \dots, n$;

(1) the forecaster chooses $I_t \in \{1, \dots, K\}$,

(2) the environment draws the reward $X_{I_t, T_{I_t}(t)}$ from ν_{I_t} and independently of the past given I_t.

At the end of the n rounds, the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Figure 1: The pure exploration problem for multi-armed bandits.

before the communication starts. The connection is then launched on the channel which is believed to be the best. Opportunistic communication systems rely on the same idea. Again the cumulative regret during the exploration phase is irrelevant since the user is only interested in the quality of its communication starting after the exploration phase.

More generally, the pure exploration problem addresses the design of strategies making the best possible use of available resources in order to optimize the performance of some decision-making task. That is, it occurs in situations with a preliminary exploration phase in which costs are not measured in terms of rewards but rather in terms of resources that come in limited budget (the number of patients in the test phase in the clinical trial setting and the time to connect in the communication example).

2 Problem setup

A stochastic multi-armed bandit game is parameterized by the number of arms K, the number of rounds (or budget) n, and K probability distributions $\nu_1, \dots, \nu_K$ associated respectively with arm $1, \dots,$ arm K. These distributions are unknown to the forecaster. For $t = 1, \dots, n$, at round t, the forecaster chooses an arm I_t in the set of arms $\{1, \dots, K\}$, and observes a reward drawn from ν_{I_t} independently from the past (actions and observations). At the end of the n rounds, the forecaster selects an arm, denoted J_n, and is evaluated in terms of the difference between the mean reward of the optimal arm and the mean reward of J_n. Precisely, let $\mu_1, \dots, \mu_K$ be the respective means of $\nu_1, \dots, \nu_K$. Let $\mu^* = \max_{k \in \{1, \dots, K\}} \mu_k$. The regret of the forecaster is

$$r_n = \mu^* - \mu_{J_n}.$$

For sake of simplicity, *we will assume that the rewards are in* $[0, 1]$ *and that there is a unique optimal arm.* Let i^* denote this arm (so, $\mu_{i^*} = \mu^*$). For $i \neq i^*$, we introduce the following suboptimality measure of arm i:

$$\Delta_i = \mu^* - \mu_i.$$

For reasons that will be obvious later, we also define Δ_{i^*} as the minimal gap

$$\Delta_{i^*} = \min_{i \neq i^*} \Delta_i.$$

We introduce the notation $(i) \in \{1, \dots, K\}$ to denote the i–th best arm (with ties break arbitrarily), hence

$$\Delta_{i^*} = \Delta_{(1)} = \Delta_{(2)} \leq \Delta_{(3)} \leq \ldots \leq \Delta_{(K)}.$$

Let e_n denote the probability of error, that is the probability that the recommendation is not the optimal one:

$$e_n = \mathbb{P}(J_n \neq i^*).$$

We have $\mathbb{E}r_n = \sum_{i \neq i^*} \mathbb{P}(J_n = i)\Delta_i$, and consequently

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

As a consequence of this equation, up to a second order term, e_n and $\mathbb{E}r_n$ behave similarly, and it does not harm to focus on the probability e_n.

For each arm i and all time rounds $t \geq 1$, we denote by $T_i(t)$ the number of times arm i was pulled from rounds 1 to t, and by $X_{i,1}, X_{i,2}, \dots, X_{i,T_i(t)}$ the sequence of associated rewards. Introduce $\widehat{X}_{i,s} =$

$\frac{1}{s}\sum_{t=1}^{s} X_{i,t}$ the empirical mean of arm i after s pulls. In the following, the symbol c will denote a positive numerical constant which may differ from line to line.

The goal of this work is to propose allocation strategies with small regret, and possibly as small as the best allocation strategy which would know beforehand the distributions $\nu_1, \dots, \nu_K$ up to a permutation. Before going further, note that the goal is unachievable for all distributions $\nu_1, \dots, \nu_K$: a policy cannot perform as well as the "oracle" allocation strategy in every particular cases. For instance, when the supports of $\nu_1, \dots, \nu_K$ are disjoint, the oracle forecaster almost surely identifies an arm by a single draw of it. As a consequence, it has almost surely zero regret for any $n \geq K$. The generic policy which does not have any knowledge on the K distributions cannot reproduce this performance for any K-tuple of disjointly supported distributions. In this work, the above goal of deciding as well as an oracle will be reached for the set of Bernoulli distributions with parameters in $(0, 1)$, but the algorithms are defined for any distributions supported in $[0, 1]$.

We would like to mention that the case $K = 2$ is unique and simple since, as we will indirectly see, it is optimally solved by the uniform allocation strategy consisting in drawing each arm $n/2$ times (up to rounding problem), and at the end recommending the arm with the highest empirical mean. Therefore, our main contributions concern more the problem of the budget allocation when $K \geq 3$. The hardness of the task will be characterized by the following quantities

$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2} \qquad \text{and} \qquad H_2 = \max_{i \in \{1,\dots,K\}} i\Delta_{(i)}^{-2}.$$

These quantities are equivalent up to a logarithmic factor since we have (see Section 6.1)

$$H_2 \leq H_1 \leq \log(2K) H_2. \tag{1}$$

Intuitively, we will show that these quantities are indeed characteristic of the hardness of the problem, in the sense that they give the order of magnitude of the number of samples required to find the best arm with a reasonable probability. This statement will be made precise in the rest of the paper, in particular through Theorem 2 and Theorem 4.

Outline. In Section 3, we propose a highly exploring policy based on upper confidence bounds, called UCB-E (Upper Confidence Bound Exploration), in the spirit of UCB1 Auer et al. (2002). We prove that this algorithm, provided that it is appropriately tuned, has an upper bound on the probability of error e_n of order $\exp\big(-c\frac{n}{H_1}\big)$. The core problem of this policy is the tuning of the parameter. The optimal value of the parameter depends on H_1, which has no reason to be known beforehand by the forecaster, and which, to our knowledge, cannot be estimated from past observations with sufficiently high confidence in order that the resulting algorithm still satisfies a similar bound on e_n.

To get round this limitation, in Section 4, we propose a simple new policy called SR (Successive Rejects) that progressively rejects the arms which seem to be suboptimal. This algorithm is parameter-free and its probability of error e_n is at most of order $\exp\big(-\frac{n}{\log(2K)H_2}\big)$. Since $H_2 \leq H_1 \leq \log(2K)H_2$, up to at most a logarithmic term in K, the algorithm performs as well as UCB-E while not requiring the knowledge of H_1.

In Section 5, we prove that H_1 and H_2 truly represent the hardness of the problem (up to a logarithmic factor). Precisely, we consider a forecaster which knows the reward distributions of the arms *up to a permutation*. When these distributions are of Bernoulli type with parameter in $[p, 1-p]$ for some $p > 0$, there exists a permutation of the distributions for which the probability of error of the (oracle) forecaster is lower bounded by $\exp\big(-\frac{cn}{p(1-p)H_2}\big)$.

Section 6 gathers some of the proofs. Section 7 provides some experiments testing the efficiency of the proposed policies and enlightening our theoretical results. We also discuss a modification of UCB-E where we perform a non-trivial online estimation of H_1. We conclude in Section 8.

Example. To put in perspective the results we just mentioned, let us consider a specific example with Bernoulli distributions. Let $\nu_1 = Ber\left(\frac{1}{2}\right)$, and $\nu_i = Ber\left(\frac{1}{2} - \frac{1}{K^i}\right)$ for $i \in \{2, \dots, K\}$. Here, one can easily check that $H_2 = 2K^{2K}$. Thus, in this case, the probability of missing the best arm of SR is at most of order $\exp\big(-\frac{n}{2\log(2K)K^{2K}}\big)$. Moreover, in Section 5, we prove that there does not exist any forecaster (even with the knowledge of the distributions up to a permutation) with a probability of missing the best arm smaller than $\exp\big(-\frac{11n}{K^{2K}}\big)$ for infinitely many n. Thus, our analysis finds that, for this particular reward distributions, the number of samples required to find the best arm is at least (of order of) K^{2K}, and SR actually finds it with (of order of) $\log(K)K^{2K}$ samples.

3 Highly exploring policy based on upper confidence bounds

In this section, we propose and study the algorithm UCB-E described in Figure 2. When a is taken of order $\log n$, the algorithm essentially corresponds to the UCB1 policy introduced in Auer et al. (2002), and

Parameter: exploration parameter $a > 0$.

For $i \in \{1, \ldots, K\}$, let $B_{i,s} = \widehat{X}_{i,s} + \sqrt{\frac{a}{s}}$ for $s \geq 1$ and $B_{i,0} = +\infty$.

For each round $t = 1, 2, \ldots, n$:
Draw $I_t \in \operatorname{argmax}_{i \in \{1,\ldots,K\}} B_{i,T_i(t-1)}$.

Let $J_n \in \operatorname{argmax}_{i \in \{1,\ldots,K\}} \widehat{X}_{i,T_i(n)}$.

Figure 2: UCB-E (Upper Confidence Bound Exploration) algorithm.

its cumulative regret is of order $\log n$. Bubeck et al. (2009) have shown that algorithms having at most logarithmic cumulative regret, have at least a (non-cumulative) regret of order $n^{-\gamma}$ for some $\gamma > 0$. So taking a of order $\log n$ is inappropriate to reach exponentially small probability of error. For our regret notion, one has to explore much more and typically use a parameter which is essentially linear in n. Precisely, we have the following result, the proof of which can be found in Section 6.2.

Theorem 1 *If UCB-E is run with parameter $0 < a \leq \frac{25}{36}\frac{n-K}{H_1}$, then it satisfies*

$$e_n \leq 2nK \exp\left(-\frac{2a}{25}\right).$$

In particular for $a = \frac{25}{36}\frac{n-K}{H_1}$, we have $e_n \leq 2nK \exp\left(-\frac{n-K}{18H_1}\right)$.

The theorem shows that the probability of error of UCB-E is at most of order $\exp(-ca)$ for $a \geq \log n$. In fact, one can easily show a corresponding lower bound. In view of this, as long as $a \leq \frac{25}{36}\frac{n-K}{H_1}$, we can essentially say: the more we explore (i.e., the larger a is), the smaller the regret is. Besides, the smallest upper bound on the probability of error is obtained for a of order n/H_1, and is therefore exponentially decreasing with n. The constant H_1 depends not only on how close the mean rewards of the two best arms are, but also on the number of arms and how close their mean reward is to the optimal mean reward. This constant should be seen as the order of the minimal number n for which the recommended arm is the optimal one with high probability. In Section 5, we will show that H_1 is indeed a good measure of the hardness of the task by showing that no forecaster satisfies $e_n \leq \exp\left(-\frac{cn}{H_2}\right)$ for any distributions $\nu_1, \ldots, \nu_K$, where we recall that H_2 satisfies $H_2 \leq H_1 \leq \log(2K)H_2$.

One interesting message to take from the proof of Theorem 1 is that, with probability at least $1 - 2nK\exp\left(-\frac{2a}{25}\right)$, the number of draws of any suboptimal arm i is of order $a\Delta_i^{-2}$. This means that the optimal arm will be played at least $n - caH_1$, showing that for too small a, UCB-E "exploits" too much in view of our regret target. Theorem 1 does not specify how the algorithm performs when a is larger than $\frac{25}{36}\frac{n-K}{H_1}$. Nevertheless, similar arguments than the ones in the proof show that for large a, with high probability, only low rewarding arms are played of order $a\Delta_i^{-2}$ times, whereas the best ones are all drawn the same number of times up to a constant factor. The number of these similarly drawn arms grows with a. In the limit, when a goes to infinity, UCB-E is exactly the uniform allocation strategy studied in Bubeck et al. (2009). In general[1], the uniform allocation has a probability of error which can be lower and upper bounded by a quantity of the form $\exp\left(-c\frac{n\Delta_{i^*}^2}{K}\right)$. It consequently performs much worse than UCB-E for $a = \frac{25}{36}\frac{n-K}{H_1}$, since $H_1 \leq K\Delta_{i^*}^{-2}$, and potentially $H_1 \ll K\Delta_{i^*}^{-2}$ for very large number of arms with heterogeneous mean rewards.

One straightforward idea to cope with the absence of an oracle telling us the value of H_1 would be to estimate online the parameter H_1 and use this estimation in the algorithm. Unfortunately, we were not able to prove, and do not believe that, this modified algorithm generally attains the expected rate of convergence. Indeed, overestimating H_1 leads to low exploring, and in the event when the optimal arm has given abnormally low rewards, the arm stops being drawn by the policy, its estimated mean reward is thus not corrected, and the arm is finally not recommended by the policy. On the contrary, underestimating H_1 leads to draw too much the suboptimal arms, precluding a sufficiently accurate estimation of the mean rewards of the best arms. For this last case, things are in fact much more subtle than what can be retranscribed in these few lines,

[1]We say "in general" to rule out some trivial cases (like when the reward distributions are all Dirac distributions) in which the probability of error e_n would be much smaller.

Let $A_1 = \{1, \ldots, K\}$, $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $n_0 = 0$ and for $k \in \{1, \ldots, K-1\}$,

$$n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil .$$

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i for $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$ (we only remove one element from A_k, if there is a tie, select randomly the arm to dismiss among the worst arms).

Let J_n be the unique element of A_K.

Figure 3: SR (Successive Rejects) algorithm.

and we notice that keeping track of a lower bound on H_1 would lead to the correct rate only under appropriate assumptions on the decrease of the sequence $\Delta_{(k)}$, $k \in \{1, \ldots, K\}$. In Section 7 we push this idea and propose a way to estimate online H_1, however we solely justify the corresponding algorithm by experiments. In the next section we propose an algorithm which does not suffer from these limitations.

4 Successive Rejects algorithm

In this section, we describe and analyze a new algorithm, SR (Successive Rejects), see Figure 3 for its precise description. Informally it proceeds as follows. First the algorithm divides the time (i.e., the n rounds) in $K-1$ phases. At the end of each phase, the algorithm dismisses the arm with the lowest empirical mean. During the next phase, it pulls equally often each arm which has not been dismissed yet. The recommended arm J_n is the last surviving arm. The length of the phases are carefully chosen to obtain an optimal (up to a logarithmic factor) convergence rate. More precisely, one arm is pulled $n_1 = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K} \right\rceil$ times, one $n_2 = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K-1} \right\rceil$ times, ..., and two arms are pulled $n_{K-1} = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{2} \right\rceil$ times. SR does not exceed the budget of n pulls, since, from the definition $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, we have

$$n_1 + \ldots + n_{K-1} + n_{K-1} \leq K + \frac{n-K}{\overline{\log}(K)} \left(\frac{1}{2} + \sum_{k=1}^{K-1} \frac{1}{K+1-k} \right) = n.$$

For $K = 2$, up to rounding effects, SR is just the uniform allocation strategy.

Theorem 2 *The probability of error of SR satisfies*

$$e_n \leq \frac{K(K-1)}{2} \exp\left(-\frac{n-K}{\overline{\log}(K) H_2} \right).$$

Proof: We can assume that the sequence of rewards for each arm is drawn before the beginning of the game. Thus the empirical reward for arm i after s pulls is well defined even if arm i has not been actually pulled s times. During phase k, at least one of the k worst arms is surviving. So, if the optimal arm i^* is dismissed at the end of phase k, it means that $\widehat{X}_{i^*,n_k} \leq \max_{i \in \{(K),(K-1),\ldots,(K+1-k)\}} \widehat{X}_{i,n_k}$. By a union bound and Hoeffding's inequality, the probability of error $e_n = \mathbb{P}(A_K \neq \{i^*\})$ thus satisfies

$$\begin{aligned}
e_n &\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}(\widehat{X}_{i^*,n_k} \leq \widehat{X}_{(i),n_k}) \\
&\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}(\widehat{X}_{(i),n_k} - \mu_{(i)} + \mu^* - \widehat{X}_{i^*,n_k} \geq \Delta_{(i)}) \\
&\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \exp\left(-n_k \Delta_{(i)}^2\right) \leq \sum_{k=1}^{K-1} k \exp\left(-n_k \Delta_{(K+1-k)}^2\right).
\end{aligned}$$

We conclude the proof by noting that by definition of n_k and H_2, we have

$$n_k \Delta_{(K+1-k)}^2 \geq \frac{n-K}{\overline{\log}(K)} \frac{1}{(K+1-k)\Delta_{(K+1-k)}^{-2}} \geq \frac{n-K}{\overline{\log}(K) H_2}. \tag{2}$$

■

The following theorem provides a deeper understanding of how SR works. It lower bounds the sampling times of the arms and shows that at the end of phase k, we have a high-confidence estimation of $\Delta_{(K+1-k)}$ up to numerical constant factor. This intuition will prove to be useful in Section 7, see in particular Figure 4.

Theorem 3 *With probability at least* $1 - \frac{K^3}{2}\exp\big(-\frac{n-K}{4\overline{\log}(K)H_2}\big)$*, for any arm* j*, we have*

$$T_j(n) \geq \frac{n-K}{4\overline{\log}(K) H_2 \Delta_j^2}. \tag{3}$$

With probability at least $1 - K^3\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$*, for any* $k \in \{1,\dots,K-1\}$*, the dismissed arm* $\ell_k = A_{k+1} \setminus A_k$ *at the end of phase* k *satisfies*

$$\frac{1}{4}\Delta_{(K+1-k)} \leq \frac{1}{2}\Delta_{\ell_k} \leq \max_{m \in A_k} \widehat{X}_{m,n_k} - \widehat{X}_{\ell_k,n_k} \leq \frac{3}{2}\Delta_{\ell_k} \leq 3\Delta_{(K+1-k)}. \tag{4}$$

Proof: We consider the event $\mathcal{E}$ on which for any $k \in \{1,\dots,K-1\}$, for any arm ℓ in the worst k arms, and any arm j such that $2\Delta_j \leq \Delta_\ell$, we have

$$\widehat{X}_{j,n_k} - \widehat{X}_{\ell,n_k} > 0.$$

This event holds with probability at least $1 - \frac{K^3}{2}\exp\big(-\frac{n-K}{4\overline{\log}(K)H_2}\big)$, since, from Hoeffding's inequality, a union bound and (2), we have

$$\begin{aligned}
&\sum_{k=1}^{K-1} \sum_{\substack{\ell \in \{(K),(K-1),\dots,(K+1-k)\} \\ j:2\Delta_j \leq \Delta_\ell}} \mathbb{P}\Big(\widehat{X}_{j,n_k} - \widehat{X}_{\ell,n_k} \leq 0\Big) \\
&\leq \sum_{k=1}^{K-1} \sum_{\substack{\ell \in \{(K),(K-1),\dots,(K+1-k)\} \\ j:2\Delta_j \leq \Delta_\ell}} \exp\big(-n_k(\Delta_\ell - \Delta_j)^2\big) \\
&\leq \sum_{k=1}^{K-1} kK \exp\big(-n_k \frac{\Delta_{(K+1-k)}^2}{4}\big) \leq \frac{K^3}{2}\exp\Big(-\frac{n-K}{4\overline{\log}(K)H_2}\Big).
\end{aligned}$$

During phase k, at least one of the k worst arms is surviving. On the event $\mathcal{E}$, this surviving arm has an empirical mean at the end of the phase which is smaller than the one of any arm j satisfying $2\Delta_j \leq \Delta_{(K+1-k)}$. So, at the end of phase k, any arm j satisfying $2\Delta_j \leq \Delta_{(K+1-k)}$ cannot be dismissed. Now, for a given arm j, we consider two cases depending whether there exists $m \in \{1,\dots,K\}$ such that $\Delta_{(m-1)} \leq 2\Delta_j \leq \Delta_{(m)}$.
First case. If no such m exists, then we have $\Delta_j^2 T_j(n) \geq \frac{1}{4}\Delta_{(K)}^2 n_1 \geq \frac{n-K}{4\overline{\log}(K)H_2}$, so that (3) holds.
Second case. If such m exists, then, from the above argument, the arm j cannot be dismissed before the end of the phase $K+2-m$ (since there exists $K+1-m$ arms ℓ such that $\Delta_\ell \geq 2\Delta_j$). From (2), we get

$$\Delta_j^2 T_j(n) \geq \Delta_j^2 n_{K+2-m} \geq \frac{\Delta_j^2}{\Delta_{(m-1)}^2} \frac{n-K}{\overline{\log}(K)H_2} \geq \frac{n-K}{4\overline{\log}(K)H_2},$$

which ends the proof of (3). We have seen that at the end of phase k, any arm j satisfying $2\Delta_j \leq \Delta_{(K+1-k)}$ cannot be dismissed. Consequently, at the end of phase k, the dismissed arm $\ell_k = A_{k+1} \setminus A_k$ satisfies the left inequality of

$$\frac{1}{2}\Delta_{(K+1-k)} \leq \Delta_{\ell_k} \leq 2\Delta_{(K+1-k)}. \tag{5}$$

Let us now prove the right inequality by contradiction. Consider k such that $2\Delta_{(K+1-k)} < \Delta_{\ell_k}$. Arm ℓ_k thus belongs to the $k-1$ worst arms. Hence, in the first $k-1$ rejects, say at the end of phase k', an arm j with $\Delta_j \leq \Delta_{(K+1-k)}$ is dismissed. From the left inequality of (5), we get $\Delta_{(K+1-k')} \leq 2\Delta_j < \Delta_{\ell_k}$.

On the event $\mathcal{E}$, we thus have $\widehat{X}_{j,n_{k'}} - \widehat{X}_{\ell_k,n_{k'}} > 0$ (since ℓ_k belongs to the k' worst arms by the previous inequality). This contradicts the fact that j is rejected at phase k'. So (5) holds.

Now let $\mathcal{E}'$ be the event on which for any arm j, and any $k \in \{1,\dots,K-1\}$ $\left|\widehat{X}_{j,n_k} - \mu_j\right| \le \frac{\Delta_{(K+1-k)}}{8}$. Using again Hoeffding's inequality, a union bound and (2), this event holds with probability at least $1 - 2K(K-1)\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$. We now work on the event $\mathcal{E}\cap\mathcal{E}'$, which holds with probability at least $1 - K^3\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$. From (5), the dismissed arm ℓ_k at the end of phase k satisfies

$$\left|\widehat{X}_{\ell_k,n_k} - \mu_{\ell_k}\right| \le \frac{\Delta_{(K+1-k)}}{8} \le \frac{\Delta_{\ell_k}}{4}.$$

Besides, we also have

$$\Big|\max_{m\in A_k}\widehat{X}_{m,n_k} - \mu_{(1)}\Big| \le \frac{\Delta_{(K+1-k)}}{8} \le \frac{\Delta_{\ell_k}}{4}.$$

Consequently, at the end of phase k, we have

$$\frac{1}{4}\Delta_{(K+1-k)} \le \frac{1}{2}\Delta_{\ell_k} \le \max_{m\in A_k}\widehat{X}_{m,n_k} - \widehat{X}_{\ell_k,n_k} \le \frac{3}{2}\Delta_{\ell_k} \le 3\Delta_{(K+1-k)}.$$

■

5 Lower bound

In this section, we provide a general and somewhat surprising lower bound. We prove that, when the reward distributions are Bernoulli distributions with variances bounded away from 0, then for any forecaster, one can permute the distributions on the arms (before the game starts) so that the probability of missing the best arm will be at least of order $\exp\big(-\frac{cn}{H_2}\big)$. Note that, in this formulation, we allow the forecaster to *know* the reward distributions up to a permutation of the indices! However, as the lower bound expresses it, whatever Bernoulli distributions with variances bounded away from 0 are considered, the quantity H_2 is a good measure of the hardness of finding the best arm.

Theorem 4 (Lower Bound) *Let $\nu_1,\dots,\nu_K$ be Bernoulli distributions with parameters in $[p,1-p]$, $p \in (0,1/2)$. For any forecaster, there exists a permutation $\sigma : \{1,\dots,K\} \to \{1,\dots,K\}$ such that the probability error of the forecaster on the bandit problem defined by $\tilde{\nu}_1 = \nu_{\sigma(1)},\dots,\tilde{\nu}_K = \nu_{\sigma(K)}$ satisfies*

$$e_n \ge \exp\left(-\frac{(5+o(1))n}{p(1-p)H_2}\right),$$

where the $o(1)$ term depends only on K, p and n and goes to 0 when n goes to infinity (see the end of the proof).

The proof of this result is quite technical. However, it is simple to explain why we can expect such a bound to hold. Assume (without loss of generality) that the arms are ordered, i.e., $\mu_1 > \mu_2 \ge \dots \ge \mu_K$, and that all rewards $X_{i,t}$, $t \in \{1,\dots,n\}$, $i \in \{1,\dots,K\}$, are drawn before the game starts. Let $i \in \{2,\dots,K\}$. If $\widehat{X}_{1,n/i} < \widehat{X}_{i,n/i} \le \widehat{X}_{j,n/i}$ for all $j \in \{2,\dots,i-1\}$, then it seems reasonable that a good forecaster should not pull arm 1 more than n/i times, and furthermore not select it as its recommendation. One can see that, the probability of the event we just described is of order of $\exp(-c(n/i)\Delta_i^2)$. Thus, with probability at least $\exp(-cn/\max_{2\le i\le K} i\Delta_i^{-2})$, the forecaster makes an error, which is exactly the lower bound we propose. However, note that this argument does not yield a reasonable proof strategy, in particular we assumed a "good" forecaster with a "reasonable" behavior. For instance, it is obvious that the proof has to permute the arms, since a forecaster could, despite all, choose arm 1 as its recommendation, which imply a probability error of 0 as soon as the best arm is in position 1.

The main idea of our proposed proof goes as follows. A bandit problem is defined by a product distribution $\nu = \nu_1 \otimes \dots \otimes \nu_K$. One can consider that at the beginning of the game, n K-tuples of rewards are sampled from this product distribution. This defines a table of nK rewards. A forecaster will explore a subpart of this table. We want to find a permutation σ of $\{1,\dots,K\}$ so that the indices of the best arm for ν and $\tilde{\nu} = \nu_{\sigma(1)} \otimes \dots \otimes \nu_{\sigma(K)}$ are different and such that the likelihood ratio of the explored part of the table of nK rewards under ν and $\tilde{\nu}$ is at least of order $\exp(-cn/H_2)$ with probability with respect to $\nu^{\otimes n}$ lower bounded by a positive numerical constant. This would imply the claimed bound. Remark that, the "likelihood cost" of moving distribution ν_i to arm j depends on both the (Kullback-Leibler) distance between the distributions ν_i and ν_j, and the number of times arm j is pulled. Thus, we have to find the right trade-off between moving

a distribution to a "close" distribution, and the fact that the target arm should not be pulled too much. To do this, we "slice" the set of indices in a non-trivial (and non-intuitive) way. This "slicing" depends only on the reward distributions, and not on the considered forecaster. Then, to put it simply, we move the less drawn arm from one slice to the less drawn arm in the next slice. Note that the preceding sentence is not well defined, since by doing this we would get a random permutation (which of course does not make sense to derive a lower bound). However, at the cost of some technical difficulties, it is possible to circumvent this issue.

To achieve the program outlined above, as already hinted, we use the Kullback-Leibler divergence, which is defined for two probability distributions ρ, ρ' on $[0,1]$ with ρ absolutely continuous with respect to ρ' as:

$$\mathrm{KL}(\rho,\rho') = \int_0^1 \log\left(\frac{d\rho}{d\rho'}(x)\right) d\rho(x) = \mathbb{E}_{X\sim\rho}\log\left(\frac{d\rho}{d\rho'}(X)\right).$$

Another quantity of particular interest for our analysis is $\widehat{\mathrm{KL}}_{i,t}(\rho,\rho') = \sum_{s=1}^t \log\left(\frac{d\rho}{d\rho'}(X_{i,s})\right)$. In particular, note that, if arm i has distribution ρ, then this quantity represents the (non re-normalized) empirical estimation of $\mathrm{KL}(\rho,\rho')$ after t pulls of arm i. Let $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ the probability and expectation signs when we integrate with respect to the distribution $\nu^{\otimes n}$. Another important property is that for any two product distributions ν, ν', which differ only on index i, and for any event A, one has:

$$\mathbb{P}_\nu(A) = \mathbb{E}_{\nu'}\mathbb{1}_A \exp\left(-\widehat{\mathrm{KL}}_{i,T_i(n)}(\nu_i',\nu_i)\right), \tag{6}$$

since we have $\prod_{s=1}^{T_{i,n}} \frac{d\nu_i}{d\nu_i'}(X_{i,s}) = \exp\big(-\widehat{\mathrm{KL}}_{i,T_i(n)}(\nu_i',\nu_i)\big)$.

Proof: First step: Notations. Without loss of generality we can assume that ν is ordered in the sense that $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$. Moreover let $L \in \{2,\ldots,K\}$ such that $H_2 = L/\Delta_L^2$, that is for all $i \in \{1,\ldots,K\}$,

$$i/\Delta_i^2 \leq L/\Delta_L^2. \tag{7}$$

We define now recursively the following sets. Let $k_1 = 1$,

$$\Sigma_1 = \left\{i : \mu_L \leq \mu_i \leq \mu_L + \frac{\Delta_L}{L^{1/2^{k_1}}}\right\},$$

and for $j > 1$,

$$\Sigma_j = \left\{i : \mu_L + \frac{\Delta_L}{L^{1/2^{k_{j-1}}}} < \mu_i \leq \mu_L + \frac{\Delta_L}{L^{1/2^{k_j}}}\right\},$$

where k_j is the smallest integer (if it exists, otherwise set $k_j = +\infty$) such that $|\Sigma_j| > 2|\Sigma_{j-1}|$. Let $\ell = \max\{j : k_j < +\infty\}$. We define now the random variables $Z_1,\ldots,Z_\ell$ corresponding to the indices of the less sampled arms of the respective slices $\Sigma_1,\ldots,\Sigma_\ell$: for $j \in \{1,\ldots,\ell\}$,

$$Z_j \in \operatorname*{argmin}_{i\in\Sigma_j} T_i(n).$$

Finally let $Z_{\ell+1} \in \operatorname{argmin}_{i\in\{1,\ldots,L\}\setminus\{J_n\}} T_i(n)$.

Second step: Controlling $T_{Z_j}(n)$, $j \in \{1,\ldots,\ell+1\}$. We first prove that for any $j \in \{1,\ldots,\ell\}$,

$$3|\Sigma_j| \geq L^{1-\frac{1}{2^{k_{j+1}-1}}}. \tag{8}$$

To do so let us note that, by definition of k_{j+1}, we have

$$\begin{aligned} 2|\Sigma_j| &\geq \left|\left\{i : \mu_L + \Delta_L/L^{1/2^{k_j}} < \mu_i \leq \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right| \\ &\geq \left|\left\{i : \mu_i \leq \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right| - (|\Sigma_1| + \ldots + |\Sigma_{j-1}|). \end{aligned}$$

Now remark that, by definition again, we have $|\Sigma_1| + \ldots + |\Sigma_{j-1}| \leq (2^{-(j-1)} + \ldots + 2^{-1})|\Sigma_j| \leq |\Sigma_j|$. Thus we obtain $3|\Sigma_j| \geq \left|\left\{i : \mu_i \leq \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right|$. We finish the proof of (8) with the following calculation, which makes use of (7). For any $v \geq 1$,

$$\begin{aligned} |\{i : \mu_i \leq \mu_L + \Delta_L/v\}| &= |\{i : \Delta_i \geq \Delta_L(1-1/v)\}| \\ &\geq \left|\left\{i : \sqrt{\frac{i}{L}}\Delta_L \geq \Delta_L(1-1/v)\right\}\right| \\ &= \left|\left\{i : i \geq L(1-1/v)^2\right\}\right| \geq L\left(1-(1-1/v)^2\right) \geq L/v. \end{aligned}$$

Now (8) directly entails (since a minimum is smaller than an average), for $j \in \{1, \dots, \ell\}$,

$$T_{Z_j}(n) \leq 3L^{\frac{1}{2^{k_{j+1}-1}}-1} \sum_{i \in \Sigma_j} T_i(n). \tag{9}$$

Besides, since $Z_{\ell+1}$ is the less drawn arm among $L-1$ arms, we trivially have

$$T_{Z_{\ell+1}}(n) \leq \frac{n}{L-1}. \tag{10}$$

Third step: A change of measure. Let $\nu' = \nu_L \otimes \nu_2 \otimes \cdots \otimes \nu_K$ be a modified product distribution where we replaced the best distribution by ν_L. Now let us consider the event

$$C_n = \Big\{ \forall t \in \{1, \dots, n\},\ i \in \{2, \dots, L\},\ j \in \{1, \dots, L\},$$
$$\widehat{\mathrm{KL}}_{i,t}(\nu_i, \nu_j) \leq t\ \mathrm{KL}(\nu_i, \nu_j) + o_n \qquad \text{and} \qquad \widehat{\mathrm{KL}}_{1,t}(\nu_L, \nu_j) \leq t\ \mathrm{KL}(\nu_L, \nu_j) + o_n \Big\},$$

where $o_n = 2\log(p^{-1})\sqrt{n\log(2L)}$. From Hoeffding's maximal inequality, one can prove that we have $\mathbb{P}_{\nu'}(C_n) \geq 1/2$. We thus have $\sum_{1 \leq z_1, \dots, z_{\ell+1} \leq L} \mathbb{P}_{\nu'}\big(C_n \cap \{Z_1 = z_1, \dots, Z_{\ell+1} = z_{\ell+1}\}\big) \geq 1/2$. Moreover note that $Z_1, \dots, Z_\ell$ are all distinct. Thus there exist $\ell+1$ constants $z_1, \dots, z_{\ell+1}$ such that, for $A_n = C_n \cap \{Z_1 = z_1, \dots, Z_{\ell+1} = z_{\ell+1}\}$, we have

$$\mathbb{P}_{\nu'}(A_n) \geq \frac{1}{2L \times L!}. \tag{11}$$

Since, by definition $Z_{\ell+1} \neq J_n$, we have

$$A_n \subset \{J_n \neq z_{\ell+1}\}. \tag{12}$$

In the following we treat differently the cases $z_{\ell+1} = 1$ and $z_{\ell+1} \neq 1$. First, let us assume that $z_{\ell+1} = 1$. Then, an application of (6) and (12) directly gives, by definition of A_n,

$$\begin{aligned}
e_n(\nu) = \mathbb{P}_\nu(J_n \neq 1) &= \mathbb{E}_{\nu'} \mathbb{1}_{J_n \neq 1} \exp\Big(- \widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L, \nu_1) \Big) \\
&\geq \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\Big(- \widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L, \nu_1) \Big) \\
&\geq \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\Big(- o_n - T_{Z_{\ell+1}}(n) \mathrm{KL}(\nu_L, \nu_1) \Big) \\
&\geq \frac{1}{2L \times L!} \exp\Big(- o_n - \frac{n}{L-1} \mathrm{KL}(\nu_L, \nu_1) \Big),
\end{aligned}$$

where we used (10) and (11) for the last equation. Now, for any $p, q \in [0,1]$, the KL divergence between Bernoulli distributions of parameters p and q satisfies

$$\mathrm{KL}(Ber(p), Ber(q)) \leq \frac{(p-q)^2}{q(1-q)}. \tag{13}$$

This can be seen by using $\log u \leq u - 1$ on the two logarithmic terms in $\mathrm{KL}(Ber(p), Ber(q))$. In particular, it implies $\mathrm{KL}(\nu_L, \nu_1) \leq \frac{\Delta_L^2}{p(1-p)}$, which concludes the proof in the case $z_{\ell+1} = 1$.

Assume now that $z_{\ell+1} \neq 1$. In this case we prove that the lower bound holds for a well defined permuted product distribution $\tilde{\nu}$ of ν. We define it as follows. Let m be the smallest $j \in \{1, \dots, \ell+1\}$ such that $z_m = z_{\ell+1}$. Now we set $\tilde{\nu}$ as follows: $\tilde{\nu}_{z_m} = \nu_1$, $\tilde{\nu}_{z_{m-1}} = \nu_{z_m}$, …, $\tilde{\nu}_{z_1} = \nu_{z_2}, \tilde{\nu}_1 = \nu_{z_1}$, and $\tilde{\nu}_j = \nu_j$ for other values of j in $\{1, \dots, K\}$. Remark that $\tilde{\nu}$ is indeed the result of a permutation of the distributions of ν. Again, an application of (6) and (12) gives, by definition of A_n,

$$\begin{aligned}
e_n(\tilde{\nu}) &= \mathbb{P}_{\tilde{\nu}}(J_n \neq z_m) \\
&= \mathbb{E}_{\nu'} \mathbb{1}_{J_n \neq z_m} \exp\Big(- \widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L, \nu_{z_1}) - \sum_{j=1}^{m-1} \widehat{\mathrm{KL}}_{z_j, T_{z_j}(n)}(\nu_{z_j}, \nu_{z_{j+1}}) - \widehat{\mathrm{KL}}_{z_m, T_{z_m}(n)}(\nu_{z_m}, \nu_{z_1}) \Big) \\
&\geq \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\Big(- (m+1) o_n - T_1(n) \mathrm{KL}(\nu_L, \nu_{Z_1}) - \sum_{j=1}^{m-1} T_{Z_j}(n) \mathrm{KL}(\nu_{Z_j}, \nu_{Z_{j+1}}) \\
&\qquad\qquad - T_{Z_m}(n) \mathrm{KL}(\nu_{Z_m}, \nu_{Z_1}) \Big).
\end{aligned} \tag{14}$$

From (13), the definition of Σ_j, and since the parameters of the Bernoulli distributions are in $[p, 1-p]$, we have $\mathrm{KL}(\nu_L, \nu_{Z_1}) \leq \frac{1}{p(1-p)} \frac{\Delta_L^2}{L}$, $\mathrm{KL}(\nu_{Z_m}, \nu_{Z_1}) \leq \frac{\Delta_L^2}{p(1-p)}$, and for any $j \in \{1, \dots, m-1\}$,

$$\mathrm{KL}(\nu_{Z_j}, \nu_{Z_{j+1}}) \leq \frac{1}{p(1-p)} \left(\frac{\Delta_L}{L^{1/2^{k_{j+1}}}} \right)^2 .$$

Reporting these inequalities, as well as (9), (10) and (11) in (14), we obtain:

$$\begin{aligned} e_n(\tilde{\nu}) &\geq \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\left(-(m+1)o_n - 3\frac{\Delta_L^2}{p(1-p)L} \left(T_1(n) + \sum_{j=1}^{m-1} \sum_{i \in \Sigma_j} T_i(n) + \frac{nL}{3(L-1)} \right) \right) \\ &\geq \frac{1}{2L \times L!} \exp\left(-L\, o_n - 3n\frac{\Delta_L^2}{p(1-p)L} \left(1 + \frac{L}{3(L-1)} \right) \right) \end{aligned}$$

Since $L \leq K$ and $2K \times K! \leq \exp\big(2K \log(K)\big)$ and from the definitions of o_n and L, we obtain

$$e_n(\tilde{\nu}) \geq \exp\left(-2K\log(K) - 2K\log(p^{-1})\sqrt{n\log(2K)} - 5\frac{n}{p(1-p)H_2} \right),$$

which concludes the proof. ■

6 Proofs

6.1 Proof of Inequalities (1)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$. Remark that $\log(K+1) - 1/2 \leq \overline{\log}(K) \leq \log(K) + 1/2 \leq \log(2K)$. Precisely, we will prove

$$H_2 \leq H_1 \leq \overline{\log}(K)\, H_2,$$

which is tight to the extent that the right inequality is an equality when for some $0 < c \leq 1/\sqrt{K}$, we have $\Delta_{(i)} = \sqrt{i}c$ for any $i \neq i^*$, and the left inequality is an equality if all Δ_i's are equal.

Proof: The left inequality follows from: for any $i \in \{1, \dots, K\}$, $H_1 = \sum_{k=1}^{K} \Delta_{(k)}^{-2} \geq \sum_{k=1}^{i} \Delta_{(i)}^{-2} \geq i\Delta_{(i)}^{-2}$. The right inequality directly comes from $\sum_{i=1}^{K} \Delta_{(i)}^{-2} = \Delta_{(2)}^{-2} + \sum_{i=2}^{K} \frac{1}{i} i\Delta_{(i)}^{-2} \leq \overline{\log}(K) \max_{i \in \{1,\dots,K\}} i\Delta_{(i)}^{-2}$.

6.2 Proof of Theorem 1

First step. Let us consider the event

$$\xi = \left\{ \forall i \in \{1, \dots, K\}, s \in \{1, \dots, n\}, |\widehat{X}_{i,s} - \mu_i| < \frac{1}{5}\sqrt{\frac{a}{s}} \right\}.$$

From Hoeffding's inequality and a union bound, we have $\mathbb{P}(\xi) \geq 1 - 2nK \exp\left(-\frac{2a}{25}\right)$. In the following, we prove that on the event ξ we have $J_n = i^*$, which concludes the proof. Since J_n is the empirical best arm, and given that we are on ξ, it is enough to prove that

$$\frac{1}{5}\sqrt{\frac{a}{T_i(n)}} \leq \frac{\Delta_i}{2}, \forall i \in \{1, \dots, K\},$$

or equivalently:

$$T_i(n) \geq \frac{4}{25}\frac{a}{\Delta_i^2}, \forall i \in \{1, \dots, K\}. \tag{15}$$

Second step. Firstly we prove by induction that

$$T_i(t) \leq \frac{36}{25}\frac{a}{\Delta_i^2} + 1, \forall i \neq i^*. \tag{16}$$

It is obviously true at time $t = 1$. Now assume that the formula is true at time $t-1$. If $I_t \neq i$ then $T_i(t) = T_i(t-1)$ and the formula still holds. On the other hand, if $I_t = i$, then in particular it means that $B_{i,T_i(t-1)} \geq B_{i^*,T_{i^*}(t-1)}$. Moreover, since we are on ξ, we have $B_{i^*,T_{i^*}(t-1)} \geq \mu^*$ and $B_{i,T_i(t-1)} \leq \mu_i + \frac{6}{5}\sqrt{\frac{a}{T_i(t-1)}}$. Thus, we have $\frac{6}{5}\sqrt{\frac{a}{T_i(t-1)}} \geq \Delta_i$. By using $T_i(t) = T_i(t-1) + 1$, we obtain (16).

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \ldots, K-1\}$, let $n_k = \left\lceil \frac{1}{\log(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$, $t_k = n_1 + \ldots n_{k-1} + (K-k+1)n_k$.
For $i \in \{1, \ldots, K\}$ and $a > 0$, let $B_{i,s}(a) = \widehat{X}_{i,s} + \sqrt{\frac{a}{s}}$ for $s \geq 1$ and $B_{i,0} = +\infty$.

Algorithm: For each phase $k = 0, 1, \ldots, K-1$:
Let $\widehat{H}_{1,k} = K$ if $k = 0$, and otherwise

$$\widehat{H}_{1,k} = \max_{K-k+1 \leq i \leq K} i \widehat{\Delta}_{<i>}^{-2},$$

where $\widehat{\Delta}_i = \left(\max_{1 \leq j \leq K} \widehat{X}_{j,T_j(t_k)}\right) - \widehat{X}_{i,T_i(t_k)}$ and $< i >$ is an ordering such that $\widehat{\Delta}_{<1>} \leq \ldots \leq \widehat{\Delta}_{<K>}$.

For $t = t_k + 1, \ldots, t_{k+1}$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} B_{i,T_i(t-1)}(cn/\widehat{H}_{1,k})$.

Recommendation: Let $J_n \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} \widehat{X}_{i,T_i(n)}$.

Figure 4: Adaptive UCB-E algorithm. Its intuitive justification goes as follows: The time points t_k correspond to the moments where the Successive Rejects algorithm would dismiss an arm. Intuitively, in light of Theorem 3, one can say that at time t_k a good algorithm should have reasonable approximation of the gaps between the best arm and the k worst arms, that is the quantities $\Delta_{(K-k+1)}, \ldots, \Delta_{(K)}$. Now with these quantities, one can build a lower estimate of H_2 and thus also of H_1. We use this estimate between the time points t_k and t_{k+1} to tune the parameter a of UCB-E.

Now we prove an other useful formula:

$$T_i(t) \geq \frac{4}{25} \min\left(\frac{a}{\Delta_i^2}, \frac{25}{36}(T_{i^*}(t) - 1)\right), \forall i \neq i^*. \tag{17}$$

With the same inductive argument as the one to get equation (16), we only need to prove that this formula holds when $I_t = i^*$. By definition of the algorithm, and since we are on ξ, when $I_t = i^*$ we have for all i:

$$\mu^* + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(t-1)}} \geq \mu_i + \frac{4}{5}\sqrt{\frac{a}{T_i(t-1)}},$$

which implies

$$T_i(t-1) \geq \frac{16}{25} \frac{a}{\left(\Delta_i + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(t-1)}}\right)^2}.$$

We then obtain (17) by using $u + v \leq 2\max(u, v)$, $T_i(t) = T_i(t-1)$ and $T_{i^*}(t-1) = T_{i^*}(t) - 1$.
Third step. Recall that we want to prove equation (15). From (17), we only have to show that

$$\frac{25}{36}(T_{i^*}(n) - 1) \geq \frac{a}{\Delta_{i^*}^2},$$

where we recall that Δ_{i^*} is the minimal gap $\Delta_{i^*} = \min_{i \neq i^*} \Delta_i$. Using equation (16) we obtain:

$$T_{i^*}(n) - 1 = n - 1 - \sum_{i \neq i^*} T_i(n) \geq n - K - \frac{36}{25} a \sum_{i \neq i^*} \Delta_i^{-2} \geq \frac{36}{25} a \Delta_{i^*}^{-2},$$

where the last inequality uses $\frac{36}{25} H_1 a \leq n - K$. This concludes the proof.

7 Experiments

We propose a few simple experiments to illustrate our theoretical analysis. As a baseline comparison we use the Hoeffding Race algorithm, see Maron and Moore (1993), and the uniform strategy, which pulls equally often each arm and recommend the arm with the highest empirical mean, see Bubeck et al. (2009) for its theoretical analysis. We consider only Bernoulli distributions, and the optimal arm always has parameter $1/2$. Each experiment corresponds to a different situation for the gaps, they are either clustered in few groups,

or distributed according to an arithmetic or geometric progression. In each experiment we choose the number of samples (almost) equal to H_1 (except for the last experiment where we run it twice, the second time with $2H_1$ samples). If our understanding of the meaning of H_1 is sound, in each experiment the strategies SR and UCB-E should be able to find the best arm with a reasonable probability (which should be roughly of the same order in each experiment). We report our results in Figure 5. The parameters for the experiments are as follows:

– Experiment 1: One group of bad arms, $K = 20$, $\mu_{2:20} = 0.4$ (meaning for any $j \in \{2, \dots, 20\}$, $\mu_j = 0.4$)
– Experiment 2: Two groups of bad arms, $K = 20$, $\mu_{2:6} = 0.42$, $\mu_{7:20} = 0.38$.
– Experiment 3: Geometric progression, $K = 4$, $\mu_i = 0.5 - (0.37)^i$, $i \in \{2, 3, 4\}$.
– Experiment 4: 6 arms divided in three groups, $K = 6$, $\mu_2 = 0.42$, $\mu_{3:4} = 0.4$, $\mu_{5:6} = 0.35$.
– Experiment 5: Arithmetic progression, $K = 15$, $\mu_i = 0.5 - 0.025i$, $i \in \{2, \dots, 15\}$.
– Experiment 6: Two good arms and a large group of bad arms, $K = 20$, $\mu_2 = 0.48$, $\mu_{3:20} = 0.37$.
– Experiment 7: Three groups of bad arms, $K = 30$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$.

The different graphics should be read as follows: Each bar represents a different algorithm and the bar's height represents the probability of error of this algorithm. The correspondence between algorithms and bars is the following:

– Bar 1: Uniform sampling strategy.
– Bar 2-4: Hoeffding Race algorithm with parameters $\delta = 0.01, 0.1, 0.3$.
– Bar 5: Successive Rejects strategy.
– Bar 6-9: UCB-E with parameter $a = cn/H_1$ where respectively $c = 1, 2, 4, 8$.
– Bar 10-14: Adaptive UCB-E (see Figure 4) with parameters $c = 1/4, 1/2, 1, 2, 4$.

8 Conclusion

This work has investigated strategies for finding the best arm in a multi-armed bandit problem. It has proposed a simple parameter-free algorithm, SR, that attains optimal guarantees up to a logarithmic term (Theorem 2 and Theorem 4). A precise understanding of both SR (Theorem 3) and a UCB policy (Theorem 1) lead us to define a new algorithm, Adaptive UCB-E. It comes without guarantee of optimal rates (see end of Section 3), but performs better than SR in practice (for $c = 1$, Adaptive UCB-E outperformed SR on all the experiments we did, even those done to make it fail). One possible explanation is that SR is too static: it does not implement more data driven arguments such as: in a phase, a surviving arm performing much worse than the other ones is still drawn until the end of the phase even if it is clear that it is the next dismissed arm.

Extensions of this work may concentrate on the following problems. (i) What is a good measure of hardness when one takes into account the (empirical) variances? Do we have a good scaling with respect to the variance with the current algorithms or do we need to modify them ? (ii) Is it possible to derive a natural anytime version of Successive Rejects (without using a doubling trick)? (iii) Is it possible to close the logarithmic gap between the lower and upper bounds? (iv) How should we modify the algorithm and the analysis if one is interested in recommending the top m actions instead of a single one?

References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. of the 20th International Conference on Algorithmic Learning Theory*, 2009.

C. Domingo, R. Gavaldà, and O. Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. *Data Mining and Knowledge Discovery*, 6(2):131–152, 2002.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

O. Maron and A. W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *NIPS*, pages 59–66, 1993.

V. Mnih, Cs. Szepesvári, and J.-Y. Audibert. Empirical bernstein stopping. In *ICML*, volume 307, pages 672–679, 2008.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

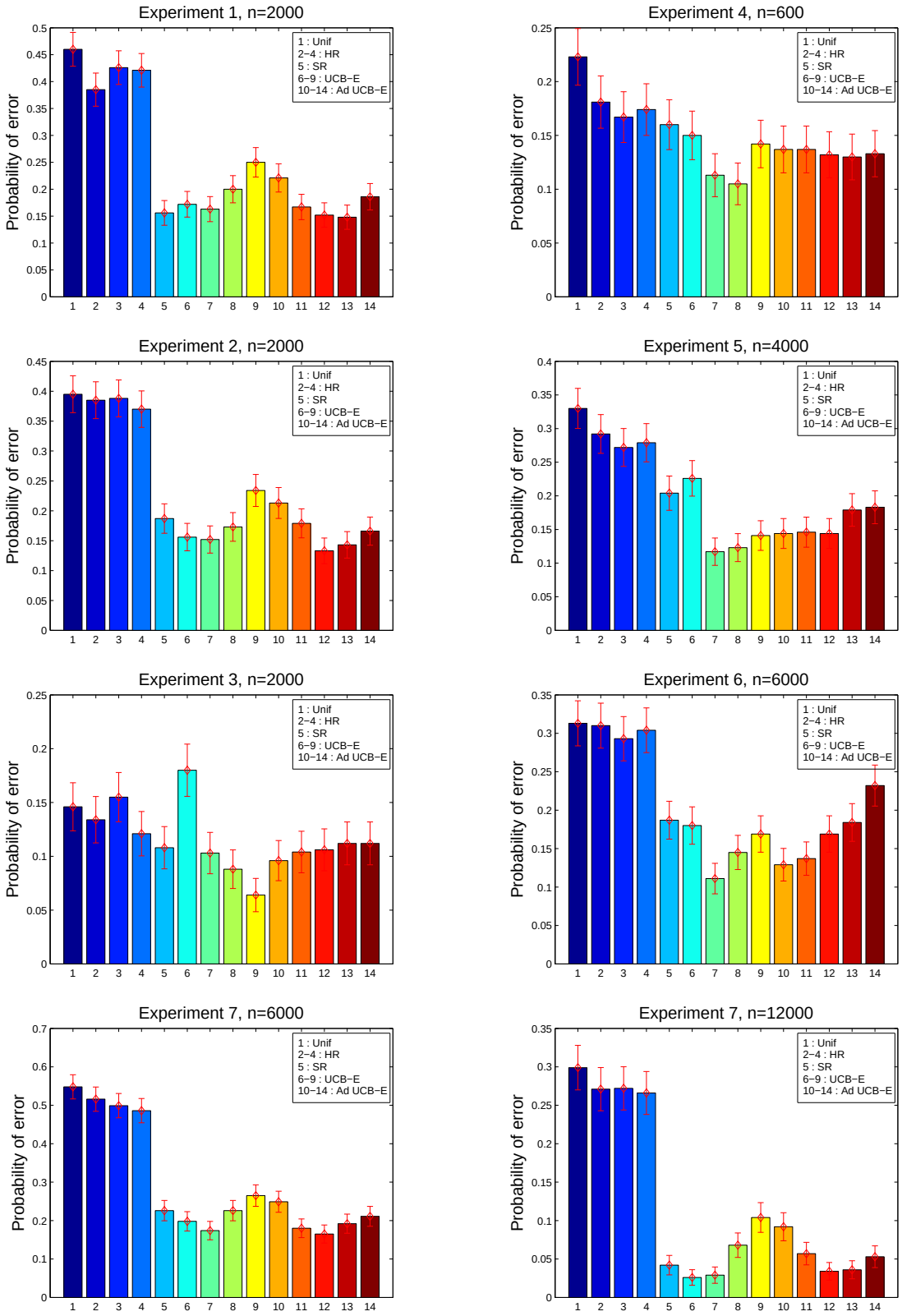

Figure 5: These results support our theoretical findings in the following sense: Despite the fact that the experiments are very different, one can see that since we use a number of samples (almost) equal to the hardness H_1, in all of them we get a probability of error of the same order, and moreover this probability is small enough to say that we identified the best arm. Note that the Successive Rejects algorithm represents in all cases a substantial improvement over both the naive uniform strategy and Hoeffding Race. These results also justify experimentally the algorithm Adaptive UCB-E. Indeed one can see that with the constant $c = 1$, we obtain better results than SR in all experiments, even in experiment 6 which was designed to be a difficult instance of Adaptive UCB-E.