Pure Exploration for Multi-Armed Bandit Problems

Sébastien Bubeck, Rémi Munos, Gilles Stoltz

Pure Exploration in Finitely–Armed and Continuously–Armed Bandits

Sébastien Bubeck*

INRIA Lille – Nord Europe, SequeL project,
40 avenue Halley, 59650 Villeneuve d'Ascq, France

Rémi Munos

INRIA Lille – Nord Europe, SequeL project,
40 avenue Halley, 59650 Villeneuve d'Ascq, France

Gilles Stoltz

Ecole Normale Supérieure, CNRS
75005 Paris, France
&
HEC Paris, CNRS,
78351 Jouy-en-Josas, France

Abstract

We consider the framework of stochastic multi-armed bandit problems and study the possibilities and limitations of forecasters that perform an on-line exploration of the arms. A forecaster is assessed in terms of its simple regret, a regret notion that captures the fact that exploration is only constrained by the number of available rounds (not necessarily known in advance), in contrast to the case when the cumulative regret is considered and when exploitation needs to be performed at the same time. We believe that this performance criterion is suited to situations when the cost of pulling an arm is expressed in terms of resources rather than rewards. We discuss the links between the simple and the cumulative regret. The main result is that the required exploration–exploitation trade-offs are qualitatively different, in view of a general lower bound on the simple regret in terms of the cumulative regret.

Key words: Multi-armed bandits, Continuously-armed bandits, Simple regret, Efficient exploration

*Corresponding author.
Email addresses: sebastien.bubeck@inria.fr (Sébastien Bubeck), remi.munos@inria.fr (Rémi Munos), gilles.stoltz@ens.fr (Gilles Stoltz)

1. Introduction

Learning processes usually face an exploration versus exploitation dilemma, since they have to get information on the environment (exploration) to be able to take good actions (exploitation). A key example is the multi-armed bandit problem [17], a sequential decision problem where, at each stage, the forecaster has to pull one out of K given stochastic arms and gets a reward drawn at random according to the distribution of the chosen arm. The usual assessment criterion of a forecaster is given by its cumulative regret, the sum of differences between the expected reward of the best arm and the obtained rewards. Typical good forecasters, like UCB of [2], trade off between exploration and exploitation.

Our setting is as follows. The forecaster may sample the arms a given number of times n (not necessarily known in advance) and is then asked to output a recommended arm. He is evaluated by his simple regret, that is, the difference between the average payoff of the best arm and the average payoff obtained by his recommendation. The distinguishing feature from the classical multi-armed bandit problem is that the exploration phase and the evaluation phase are separated. 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. 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 (the associated reward would equal a large negative value). 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. (Here, several formulæ for a cream are considered and some quantitative measurement, like skin moisturization, is performed.)

The pure exploration problem addresses the design of strategies making the best possible use of available numerical resources (e.g., as CPU time) 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.

A motivating example concerns recent works on computer-go (e.g., the MoGo program of [9]). A given time, i.e., a given amount of CPU times is given to the player to explore the possible outcome of sequences of plays and output a final decision. An efficient exploration of the search space is obtained by considering a hierarchy of forecasters minimizing some cumulative regret – see, for instance, the UCT strategy of [12] and the BAST strategy of [6]. However, the cumulative regret does not seem to be the right way to base the strategies on, since the simulation costs are the same for exploring all options, bad and good ones. This observation was actually the starting point of the notion of simple regret and of this work.

A final related example is the maximization of some function f, observed with noise, see, e.g., [11, 5]. Whenever evaluating f at a point is costly (e.g.,

in terms of numerical or financial costs), the issue is to choose as adequately as possible where to query the value of this function in order to have a good approximation to the maximum. The pure exploration problem considered here addresses exactly the design of adaptive exploration strategies making the best use of available resources in order to make the most precise prediction once all resources are consumed.

As a remark, it also turns out that in all examples considered above, we may impose the further restriction that the forecaster ignores ahead of time the amount of available resources (time, budget, or the number of patients to be included) – that is, we seek for anytime performance.

We end this introduction with an overview of the literature. The problem of pure exploration presented above was referred to as "budgeted multi-armed bandit problem" in the open problem [14] (where, however, another notion of regret than simple regret is considered). [18] solves the pure exploration problem in a minmax sense for the case of two arms only and rewards given by probability distributions over $[0, 1]$. [8] and [15] consider a related setting where forecasters perform exploration during a random number of rounds T and aim at identifying an ε–best arm. They study the possibilities and limitations of policies achieving this goal with overwhelming $1 - \delta$ probability and indicate in particular upper and lower bounds on (the expectation of) T. Another related problem is the identification of the best arm (with high probability). However, this binary assessment criterion (the forecaster is either right or wrong in recommending an arm) does not capture the possible closeness in performance of the recommended arm compared to the optimal one, which the simple regret does. Moreover unlike the latter, this criterion is not suited for a distribution-free analysis.

2. Problem setup, notation, structure of the paper

We consider a sequential decision problem given by stochastic multi-armed bandits. $K \geqslant 2$ arms, denoted by $i = 1, \ldots, K$, are available and the i–th of them is parameterized by a fixed (unknown) probability distribution ν_i over $[0, 1]$, with expectation denoted by μ_i. At those rounds when it is pulled, its associated reward is drawn at random according to ν_i, independently of all previous rewards. For each arm i and all time rounds $n \geqslant 1$, we denote by $T_i(n)$ the number of times arm i was pulled from rounds 1 to n, and by $X_{i,1}, X_{i,2}, \ldots, X_{i,T_i(n)}$ the sequence of associated rewards.

The forecaster has to deal simultaneously with two tasks, a primary one and an associated one.

The associated task consists in exploration, i.e., the forecaster should indicate at each round t the arm I_t to be pulled, based on past rewards (so that I_t is a random variable). Then the forecaster gets to see the associated reward Y_t, also denoted by $X_{I_t,T_{I_t}(t)}$ with the notation above. The sequence of random variables (I_t) is referred to as an allocation strategy.

The primary task is to output at the end of each round t a recommendation J_t to be used in a one-shot instance if/when the environment sends some stopping

Parameters: K probability distributions for the rewards of the arms, $\nu_1, \ldots, \nu_K$.

For each round $t = 1, 2, \ldots$,

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

(2) the environment draws the reward Y_t for that action (also denoted by $X_{I_t, T_{I_t}(t)}$ with the notation introduced in the text);

(3) the forecaster outputs a recommendation $J_t \in \{1, \ldots, K\}$;

(4) if the environment sends a stopping signal, then the game takes an end; otherwise, the next round starts.

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

signal meaning that the exploration phase is over. The sequence of random variables (J_t) is referred to as a recommendation strategy. In total, a forecaster is given by an allocation and a recommendation strategy.

Figure 1 summarizes the description of the sequential game and points out that the information available to the forecaster for choosing I_t, respectively J_t, is formed by the $X_{i,s}$ for $i = 1, \ldots, K$ and $s = 1, \ldots, T_i(t-1)$, respectively, $s = 1, \ldots, T_i(t)$. Note that we also allow the forecaster to use an external randomization in the definition of I_t and J_t.

As we are only interested in the performances of the recommendation strategy (J_t), we call this problem the pure exploration problem for multi-armed bandits and evaluate the forecaster through its simple regret, defined as follows. First, we denote by

$$\mu^* = \mu_{i^*} = \max_{i=1,\ldots,K} \mu_i$$

the expectation of the rewards of the best arm i^* (a best arm, if there are several of them with same maximal expectation). A useful notation in the sequel is the gap $\Delta_i = \mu^* - \mu_i$ between the maximal expected reward and the one of the i–th arm; as well as the minimal gap

$$\Delta = \min_{i:\Delta_i > 0} \Delta_i \ .$$

Now, the simple regret at round n equals the regret on a one-shot instance of the game for the recommended arm J_n, that is, put more formally,

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

A quantity of related interest is the cumulative regret at round n, which is defined as

$$R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t} \ .$$

A popular treatment of the multi-armed bandit problems is to construct forecasters ensuring that $\mathbb{E}R_n = o(n)$, see, e.g., [13] or [2], and even $R_n = o(n)$ a.s.,

as follows, e.g., from [3, Theorem 6.3] together with a martingale argument. The quantity $r'_t = \mu^* - \mu_{I_t}$ is sometimes called instantaneous regret. It differs from the simple regret r_t and in particular, $R_n = r'_1 + \ldots + r'_n$ is in general not equal to $r_1 + \ldots + r_n$. Theorem 1, among others, will however indicate some connections between r_n and R_n.

Goal and structure of the paper. We study the links between the simple and the cumulative regret. Intuitively, an efficient allocation strategy for the simple regret should rely on some exploration–exploitation trade-off. Our main contribution (Theorem 1, Section 3) is a lower bound on the simple regret in terms of the cumulative regret suffered in the exploration phase, showing that the trade-off involved in the minimization of the simple regret is somewhat different from the one for the cumulative regret. In particular it implies that the uniform allocation is a good benchmark when n is large. In Sections 4 and 5, we show how, despite all, one can fight against this negative result. For instance, some strategies designed for the cumulative regret can outperform (for moderate values of n) strategies with exponential rates of convergence for their simple regret. Finally in Section 6 we investigate the continuously-armed bandit problem where the set of arms is a topological space. In this setting we use the simple regret as a tool to characterize the spaces for which it is possible to have a sublinear cumulative regret.

3. The smaller the cumulative regret, the larger the simple regret

It is immediate that for well-chosen recommendation strategies, the simple regret can be upper bounded in terms of the cumulative regret. For instance, the strategy that at time n recommends arm i with probability $T_i(n)/n$ (recall that we allow the forecaster to use an external randomization) ensures that the simple regret satisfies $\mathbb{E}r_n = \mathbb{E}R_n/n$. Therefore, upper bounds on $\mathbb{E}R_n$ lead to upper bounds on $\mathbb{E}r_n$.

We show here that, conversely, upper bounds on $\mathbb{E}R_n$ also lead to lower bounds on $\mathbb{E}r_n$: the smaller the guaranteed upper bound on $\mathbb{E}R_n$, the larger the lower bound on $\mathbb{E}r_n$, no matter what the recommendation strategy is.

This is interpreted as a variation of the "classical" trade-off between exploration and exploitation. Here, while the recommendation strategy (J_n) relies only on the exploitation of the results of the preliminary exploration phase, the design of the allocation strategy (I_t) consists in an efficient exploration of the arms. To guarantee this efficient exploration, past payoffs of the arms have to be considered and thus, even in the exploration phase, some exploitation is needed. Theorem 1 and its corollaries aim at quantifying the needed respective amount of exploration and exploitation. In particular, to have an asymptotic optimal rate of decrease for the simple regret, each arm should be sampled a linear number of times, while for the cumulative regret, it is known that the forecaster should not do so more than a logarithmic number of times on the suboptimal arms.

Formally, our main result is as follows. It is strong in the sense that we get lower bounds for *all* possible *sets* of Bernoulli distributions $\{\nu_1, \dots, \nu_K\}$ over the rewards. Note that the stated result requires in particular that there is a unique best arm.

Theorem 1 (Main result). *For any forecaster (i.e., for any pair of allocation and recommendation strategies) and any function $\varepsilon : \{1, 2, \dots\} \to \mathbb{R}$ such that*

> *for all (Bernoulli) distributions $\nu_1, \dots, \nu_K$ on the rewards, there exists a constant $C \geqslant 0$ with $\mathbb{E}R_n \leqslant C\,\varepsilon(n)$,*

the following holds true:

> *for all sets of $K \geqslant 3$ distinct Bernoulli distributions on the rewards, with parameters different from 1, there exists a constant $D \geqslant 0$ and an ordering $\nu_1, \dots, \nu_K$ of the considered distributions such that*
>
> $$\mathbb{E}r_n \geqslant \frac{\Delta}{2}\, e^{-D\varepsilon(n)} \ .$$

Corollary 1 (General distribution-dependent lower bound). *For any forecaster, and any set of $K \geqslant 3$ distinct, Bernoulli distributions on the rewards, with parameters different from 1, there exist two constants $\beta > 0$ and $\gamma \geqslant 0$ such that, up to the choice of a good ordering of the considered distributions,*

$$\mathbb{E}r_n \geqslant \beta\, e^{-\gamma n} \ .$$

Theorem 1 is proved below and Corollary 1 follows from the fact that the cumulative regret is always bounded by n. To get further the point of the theorem, one should keep in mind that the typical (distribution-dependent) rate of growth of the cumulative regret of good algorithms, e.g., UCB1 of [2], is $\varepsilon(n) = \ln n$. This, as asserted in [13], is the optimal rate. But a recommendation strategy based on such allocation strategy is bound to suffer a simple regret that decreases at best polynomially fast. We state this result for the slight modification UCB(α) of UCB1 stated in Figure 2 and introduced by [1]; its proof relies on noting that it achieves a cumulative regret bounded by a large enough distribution-dependent constant times $\varepsilon(n) = \alpha \ln n$.

Corollary 2 (Distribution-dependent lower bound for UCB(α)). *The allocation strategy (I_t) given by the forecaster UCB(α) of Figure 2 ensures that for any recommendation strategy (J_t) and all sets of $K \geqslant 3$ distinct, Bernoulli distributions on the rewards, with parameters different from 1, there exist two constants $\beta > 0$ and $\gamma \geqslant 0$ (independent of α) such that, up to the choice of a good ordering of the considered distributions,*

$$\mathbb{E}r_n \geqslant \beta\, n^{-\gamma\alpha} \ .$$

Proof. The intuitive version of the proof of Theorem 1 is as follows. The basic idea is to consider a tie case when the best and worst arms have zero empirical

means; it happens often enough (with a probability at least exponential in the number of times we pulled these arms) and results in the forecaster basically having to pick another arm and suffering some regret. Permutations are used to control the case of untypical or naive forecasters that would despite all pull an arm with zero empirical mean, since they force a situation when those forecasters choose the worst arm instead of the best one.

Formally, we fix the forecaster (a pair of allocation and recommendation strategies) and a corresponding function ε such that the assumption of the theorem is satisfied. We denote by $\boldsymbol{p}_n = (p_{1,n}, \ldots, p_{K,n})$ the probability distribution from which J_n is drawn at random thanks to an auxiliary distribution. Note that $\boldsymbol{p}_n$ is a random vector which depends on $I_1, \ldots, I_n$ as well as on the obtained rewards $Y_1, \ldots, Y_n$. We consider below a set of $K \geqslant 3$ distinct Bernoulli distributions, satisfying the conditions of the theorem; actually, we only use below that their parameters are (up to a first ordering) such that $1 > \mu_1 > \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_K \geqslant 0$ and $\mu_2 > \mu_K$ (thus, $\mu_2 > 0$).

Step 0 introduces another layer of notation. The latter depends on permutations σ of $\{1, \ldots, K\}$. To have a gentle start, we first describe the notation when the permutation is the identity, $\sigma = \mathrm{id}$. We denote by $\mathbb{P}$ and $\mathbb{E}$ the probability and expectation with respect to the original K-tuple $\nu_1, \ldots, \nu_K$ of distributions over the arms. For $i = 1$ (respectively, $i = K$), we denote by $\mathbb{P}_{i,\mathrm{id}}$ and $\mathbb{E}_{i,\mathrm{id}}$ the probability and expectation with respect to the K-tuples formed by $\delta_0, \nu_2, \ldots, \nu_K$ (respectively, $\delta_0, \nu_2, \ldots, \nu_{K-1}, \delta_0$), where δ_0 denotes the Dirac measure on 0.

For a given permutation σ, we consider a similar notation up to a reordering, as follows. $\mathbb{P}_\sigma$ and $\mathbb{E}_\sigma$ refer to the probability and expectation with respect to the K-tuple of distributions over the arms formed by the $\nu_{\sigma^{-1}(1)}, \ldots, \nu_{\sigma^{-1}(K)}$. Note in particular that the i–th best arm is located in the $\sigma(i)$–th position. Now, we denote for $i = 1$ (respectively, $i = K$) by $\mathbb{P}_{i,\sigma}$ and $\mathbb{E}_{i,\sigma}$ the probability and expectation with respect to the K-tuple formed by the $\nu_{\sigma^{-1}(i)}$, except that we replaced the best of them, located in the $\sigma(1)$–th position, by a Dirac measure on 0 (respectively, the best and worst of them, located in the $\sigma(1)$–th and $\sigma(K)$–th positions, by Dirac measures on 0). We provide now a proof in six steps.

Step 1 lower bounds the quantity of interest by an average the maximum of the simple regrets obtained by reordering,

$$\max_\sigma \ \mathbb{E}_\sigma r_n \geqslant \frac{1}{K!} \sum_\sigma \ \mathbb{E}_\sigma r_n \geqslant \frac{\mu_1 - \mu_2}{K!} \ \sum_\sigma \mathbb{E}_\sigma \big[1 - p_{\sigma(1),n}\big] \ ,$$

where we used that under $\mathbb{P}_\sigma$, the index of the best arm is $\sigma(1)$ and the minimal regret for playing any other arm is at least $\mu_1 - \mu_2$.

Step 2 rewrites each term of the sum over σ as the product of three simple terms. We use first that $\mathbb{P}_{1,\sigma}$ is the same as $\mathbb{P}_\sigma$, except that it ensures that arm $\sigma(1)$ has zero reward throughout. Denoting by

$$C_{i,n} = \sum_{t=1}^{T_i(n)} X_{i,t}$$

the cumulative reward of the i–th arm till round n, one then gets

$$\begin{aligned}\mathbb{E}_\sigma\big[1-p_{\sigma(1),n}\big] &\geqslant \mathbb{E}_\sigma\Big[\big(1-p_{\sigma(1),n}\big)\,\mathbb{1}_{\{C_{\sigma(1),n}=0\}}\Big]\\ &= \mathbb{E}_\sigma\Big[1-p_{\sigma(1),n}\;\Big|\;C_{\sigma(1),n}=0\Big]\times\mathbb{P}_\sigma\big\{C_{\sigma(1),n}=0\big\}\\ &= \mathbb{E}_{1,\sigma}\big[1-p_{\sigma(1),n}\big]\;\mathbb{P}_\sigma\big\{C_{\sigma(1),n}=0\big\}\;.\end{aligned}$$

Second, iterating the argument from $\mathbb{P}_{1,\sigma}$ to $\mathbb{P}_{K,\sigma}$,

$$\begin{aligned}\mathbb{E}_{1,\sigma}\big[1-p_{\sigma(1),n}\big] &\geqslant \mathbb{E}_{1,\sigma}\Big[1-p_{\sigma(1),n}\;\Big|\;C_{\sigma(K),n}=0\Big]\;\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n}=0\big\}\\ &= \mathbb{E}_{K,\sigma}\big[1-p_{\sigma(1),n}\big]\;\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n}=0\big\}\end{aligned}$$

and therefore,

$$\mathbb{E}_\sigma\big[1-p_{\sigma(1),n}\big]\geqslant\mathbb{E}_{K,\sigma}\big[1-p_{\sigma(1),n}\big]\;\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n}=0\big\}\;\mathbb{P}_\sigma\big\{C_{\sigma(1),n}=0\big\}\;. \quad (1)$$

Step 3 deals with the second term in the right-hand side of (1),

$$\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n}=0\big\}=\mathbb{E}_{1,\sigma}\Big[(1-\mu_K)^{T_{\sigma(K)}(n)}\Big]\geqslant(1-\mu_K)^{\mathbb{E}_{1,\sigma}T_{\sigma(K)}(n)}\;,$$

where the equality can be seen by conditioning on $I_1,\ldots,I_n$ and then taking the expectation, whereas the inequality is a consequence of Jensen's inequality. Now, the expected number of times the suboptimal arm $\sigma(K)$ is pulled under $\mathbb{P}_{1,\sigma}$ (for which $\sigma(2)$ is the optimal arm) is bounded by the regret, by the very definition of the latter: $(\mu_2-\mu_K)\,\mathbb{E}_{1,\sigma}T_{\sigma(K)}(n)\leqslant\mathbb{E}_{1,\sigma}R_n$. Since by hypothesis (and by taking the maximum of $K!$ values), there exists a constant C such that for all σ, $\mathbb{E}_{1,\sigma}R_n\leqslant C\,\varepsilon(n)$, we finally get

$$\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n}=0\big\}\geqslant(1-\mu_K)^{C\varepsilon(n)/(\mu_2-\mu_K)}\;.$$

Step 4 lower bounds the third term in the right-hand side of (1) as

$$\mathbb{P}_\sigma\big\{C_{\sigma(1),n}=0\big\}\geqslant(1-\mu_1)^{C\varepsilon(n)/\mu_2}\;.$$

We denote by $W_n=(I_1,Y_1,\ldots,I_n,Y_n)$ the history of pulled arms and obtained payoffs up to time n. What follows is reminiscent of the techniques used in [15]. We are interested in realizations $w_n=(i_1,y_1,\ldots,i_n,y_n)$ of the history such that whenever $\sigma(1)$ was played, it got a null reward. (We denote above by $t_j(t)$ is the realization of $T_j(t)$ corresponding to w_n, for all j and t.) The likelihood of such a w_n under $\mathbb{P}_\sigma$ is $(1-\mu_1)^{t_{\sigma(1)}(n)}$ times the one under $\mathbb{P}_{1,\sigma}$. Thus,

$$\begin{aligned}\mathbb{P}_\sigma\big\{C_{\sigma(1),n}=0\big\}&=\sum\mathbb{P}_\sigma\{W_n=w_n\}\\ &=\sum(1-\mu_1)^{t_{\sigma(1)}(n)}\;\mathbb{P}_{1,\sigma}\{W_n=w_n\}=\mathbb{E}_{1,\sigma}\Big[(1-\mu_1)^{T_{\sigma(1)}(n)}\Big]\;,\end{aligned}$$

where the sums are over those histories w_n such that the realizations of the payoffs obtained by the arm $\sigma(1)$ equal $x_{\sigma(1),s} = 0$ for all $s = 1, \ldots, t_{\sigma(1)}(n)$. The argument is concluded as before, first by Jensen's inequality and then, by using that $\mu_2 \, \mathbb{E}_{1,\sigma} T_{\sigma(1)}(n) \leqslant \mathbb{E}_{1,\sigma} R_n \leqslant C \, \varepsilon(n)$ by definition of the regret and the hypothesis put on its control.

Step 5 resorts to a symmetry argument to show that as far as the first term of the right-hand side of (1) is concerned,

$$\sum_{\sigma} \mathbb{E}_{K,\sigma} \Big[1 - p_{\sigma(1),n} \Big] \geqslant \frac{K!}{2}.$$

Since $\mathbb{P}_{K,\sigma}$ only depends on $\sigma(2), \ldots, \sigma(K-1)$, we denote by $\mathbb{P}^{\sigma(2),\ldots,\sigma(K-1)}$ the common value of these probability distributions when $\sigma(1)$ and $\sigma(K)$ vary (and a similar notation for the associated expectation). We can thus group the permutations σ two by two according to these $(K-2)$–tuples, one of the two permutations being defined by $\sigma(1)$ equal to one of the two elements of $\{1, \ldots, K\}$ not present in the $(K-2)$–tuple, and the other one being such that $\sigma(1)$ equals the other such element. Formally,

$$\begin{aligned} \sum_{\sigma} \mathbb{E}_{K,\sigma} p_{\sigma(1),n} &= \sum_{j_2,\ldots,j_{K-1}} \mathbb{E}^{j_2,\ldots,j_{K-1}} \left[\sum_{j \in \{1,\ldots,K\} \setminus \{j_2,\ldots,j_{K-1}\}} p_{j,n} \right] \\ &\leqslant \sum_{j_2,\ldots,j_{K-1}} \mathbb{E}^{j_2,\ldots,j_{K-1}} \big[1 \big] = \frac{K!}{2} \, , \end{aligned}$$

where the summations over $j_2, \ldots, j_{K-1}$ are over all possible $(K-2)$–tuples of distinct elements in $\{1, \ldots, K\}$.

Step 6 simply puts all pieces together and lower bounds $\max_{\sigma} \mathbb{E}_{\sigma} r_n$ by

$$\begin{aligned} & \frac{\mu_1 - \mu_2}{K!} \sum_{\sigma} \mathbb{E}_{K,\sigma} \big[1 - p_{\sigma(1),n} \big] \, \mathbb{P}_{\sigma} \left\{ C_{\sigma(1),n} = 0 \right\} \, \mathbb{P}_{1,\sigma} \left\{ C_{\sigma(K),n} = 0 \right\} \\ \geqslant \; & \frac{\mu_1 - \mu_2}{2} \left((1 - \mu_K)^{C/(\mu_2 - \mu_K)} \, (1 - \mu_1)^{C/\mu_2} \right)^{\varepsilon(n)} \, . \end{aligned}$$

4. Upper bounds on the simple regret

In this section, we aim at qualifying the implications of Theorem 1 by pointing out that is should be interpreted as a result for large n only. For moderate values of n, strategies not pulling each arm a linear number of times in the exploration phase can have a smaller simple regret.

To do so, we consider only two natural and well-used allocation strategies. The first one is the uniform allocation, which we use as a simple benchmark; it pulls each arm a linear number of times (see Figure 2 for its formal description). The second one is UCB(α) (a variant of UCB1 introduced in [1] using an

exploration rate parameter $\alpha > 1$ and described also in Figure 2). It is designed for the classical exploration–exploitation dilemma (i.e., it minimizes the cumulative regret) and pulls suboptimal arms a logarithmic number of times only. Of course, fancier allocation strategies should also be considered in a second time but since the aim of this paper is to study the links between the cumulative and simple regret, we restrict our attention to the two discussed above.

Uniform allocation (Unif) — Plays all arms one after the other

For each round $t = 1, 2, \ldots,$

pull $I_t = [t \bmod K]$, where $[t \bmod K]$ denotes the value of t modulo K.

UCB(α) — Plays at each round the arm with the highest upper confidence bound

Parameter: exploration factor $\alpha > 1$

For each round $t = 1, 2, \ldots,$

(1) for each $i \in \{1, \ldots, K\}$, if $T_i(t-1) = 0$ let $B_{i,t} = +\infty$; otherwise, let

$$B_{i,t} = \widehat{\mu}_{i,t-1} + \sqrt{\frac{\alpha \ln t}{T_i(t-1)}} \qquad \text{where} \qquad \widehat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s} \ ;$$

(2) Pull $I_t \in \underset{i=1,\ldots,K}{\operatorname{argmax}} B_{i,t}$

(ties broken by choosing, for instance, the arm with smallest index).

Figure 2: Two allocation strategies.

In addition to these allocation strategies we consider three recommendation strategies, the ones that recommend respectively the empirical distribution of plays, the empirical best arm, or the most played arm. They are formally defined in Figure 3.

Table 1 summarizes the distribution-dependent and distribution-free bounds we could prove in this paper (the difference between the two families of bounds is whether the constants in the bounds can depend or not on the unknown distributions ν_j). It shows that two interesting couples of strategies are, on the one hand, the uniform allocation together with the choice of the empirical best arm, and on the other hand, UCB(α) together with the choice of the most played arm. The first pair was perhaps expected, the second one might be considered more surprising.

Table 1 also indicates that while for distribution-dependent bounds, the asymptotic optimal rate of decrease for the simple regret in the number n of rounds is exponential, for distribution-free bounds, this rate worsens to $1/\sqrt{n}$. A similar situation arises for the cumulative regret, see [13] (optimal $\ln n$ rate for distribution-dependent bounds) versus [3] (optimal $\sqrt{n}$ rate for distribution-free bounds).

Parameters: the history $I_1, \ldots, I_n$ of played actions and of their associated rewards $Y_1, \ldots, Y_n$, grouped according to the arms as $X_{i,1}, \ldots, X_{i,T_i(n)}$, for $i = 1, \ldots, n$

Empirical distribution of plays (EDP)
Recommends arm i with probability $T_i(n)/n$, that is, draws J_n at random according to

$$\boldsymbol{p}_n = \left(\frac{T_1(n)}{n}, \ldots, \frac{T_K(n)}{n} \right) .$$

Empirical best arm (EBA)
Only considers arms i with $T_i(n) \geqslant 1$, computes their associated empirical means

$$\widehat{\mu}_{i,n} = \frac{1}{T_i(n)} \sum_{s=1}^{T_i(n)} X_{i,s} ,$$

and forms the recommendation

$$J_n \in \operatorname*{argmax}_{i=1,\ldots,K} \widehat{\mu}_{i,n}$$

(ties broken in some way).

Most played arm (MPA)
Recommends the most played arm,

$$J_n \in \operatorname*{argmax}_{i=1,\ldots,K} T_i(n)$$

(ties broken in some way).

Figure 3: Three recommendation strategies.

Remark 1. The distribution-free lower bound in Table 1 follows from a straightforward adaptation of the proof of the lower bound on the cumulative regret in [3]; one can prove that, for $n \geqslant K \geqslant 2$,

$$\inf \sup \mathbb{E} r_n \geqslant \frac{1}{20} \sqrt{\frac{K}{n}} ,$$

where the infimum is taken over all forecasters while the supremum considers all sets of K distributions over $[0, 1]$.

4.1. A simple benchmark: the uniform allocation strategy

As explained above, the combination of the uniform allocation with the recommendation indicating the empirical best arm, forms an important theoretical benchmark. This section studies briefly its theoretical properties: the rate of decrease of its simple regret is exponential in a distribution-dependent sense and

	Distribution-dependent		
	EDP	EBA	MPA
Uniform		$\bigcirc\, e^{-\bigcirc n}$ (Pr.1)	
UCB(α)	$\bigcirc(\alpha \ln n)/n$ (Rk.2)	$\bigcirc\, n^{-\bigcirc}$ (Rk.3)	$\bigcirc\, n^{2(1-\alpha)}$ (Th.2)
Lower bound		$\bigcirc\, e^{-\bigcirc n}$ (Cor.1)	
	Distribution-free		
	EDP	EBA	MPA
Uniform		$\square\sqrt{\frac{K \ln K}{n}}$ (Cor.3)	
UCB(α)	$\square\sqrt{\frac{\alpha K \ln n}{n}}$ (Rk.2)	$\frac{\square}{\sqrt{\ln n}}$ (Rk.3)	$\square\sqrt{\frac{\alpha K \ln n}{n}}$ (Th.3)
Lower bound		$\square\sqrt{\frac{K}{n}}$ (Rk.1)	

Table 1: Distribution-dependent (top) and distribution-free (bottom) upper bounds on the expected simple regret of the considered pairs of allocation (rows) and recommendation (columns) strategies. Lower bounds are also indicated. The $\square$ symbols denote the universal constants, whereas the $\bigcirc$ are distribution-dependent constants. In parentheses, we provide the reference within this paper (index of the proposition, theorem, remark, corollary) where the stated bound is proved.

equals the optimal (up to a logarithmic term) $1/\sqrt{n}$ rate in the distribution-free case.

Below, we mean by the recommendation given by the empirical best arm at round $K\lfloor n/K \rfloor$ the recommendation $J_{K\lfloor n/K \rfloor}$ of EBA (see Figure 3), where $\lfloor x \rfloor$ denotes the lower integer part of a real number x. The reason why at round n we prefer $J_{K\lfloor n/K \rfloor}$ to J_n is only technical. The analysis is indeed simpler when all averages over the rewards obtained by each arm are over the same number of terms. This happens at rounds n multiple of K and this is why we prefer taking the recommendation of round $K\lfloor n/K \rfloor$ instead of the one of round n.

We propose first two distribution-dependent bounds, the first one is sharper in the case when there are few arms, while the second one is suited for large K.

Proposition 1 (Distribution-dependent; Unif and EBA). *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round $K\lfloor n/K \rfloor$) ensures that*

$$\mathbb{E}r_n \leqslant \sum_{i:\Delta_i>0} \Delta_i \, e^{-\Delta_i^2 \lfloor n/K \rfloor/2} \qquad \text{for all } n \geqslant K \ ;$$

and also,

$$\mathbb{E}r_n \leqslant \left(\max_{i=1,\dots,K} \Delta_i\right) \exp\left(-\frac{1}{8}\left\lfloor \frac{n}{K} \right\rfloor \Delta^2\right) \qquad \text{for all } n \geqslant \left(1 + \frac{8 \ln K}{\Delta^2}\right) K \ .$$

PROOF. To prove the first inequality, we relate the simple regret to the probability of choosing a non-optimal arm,

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} = \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{P}\{J_n = i\} \leqslant \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{P}\big\{\widehat{\mu}_{i,n} \geqslant \widehat{\mu}_{i^*,n}\big\}$$

where the upper bound follows from the fact that to be the empirical best arm, an arm i must have performed, in particular, better than a best arm i^*. We now apply Hoeffding's inequality (for i.i.d. random variables, see [10]). $\widehat{\mu}_{i,n} - \widehat{\mu}_{i^*,n}$ is an average of $\lfloor n/K \rfloor$ i.i.d. random variables bounded between -1 and 1 and with common expectation $-\Delta_i$. Thus, the probability of interest is bounded by

$$\begin{aligned}\mathbb{P}\big\{\widehat{\mu}_{i,n} - \widehat{\mu}_{i^*,n} \geqslant 0\big\} &= \mathbb{P}\Big\{\big(\widehat{\mu}_{i,n} - \widehat{\mu}_{i^*,n}\big) - \big(-\Delta_i\big) \geqslant \Delta_i\Big\} \\ &\leqslant \exp\left(-\frac{2\lfloor n/K \rfloor^2 \Delta_i^2}{4\lfloor n/K \rfloor}\right) = \exp\left(-\frac{1}{2}\left\lfloor \frac{n}{K} \right\rfloor \Delta_i^2\right) \ ,\end{aligned}$$

which yields the first result.

The second inequality is proved by resorting to a sharper concentration argument, namely, the method of bounded differences, see [16], see also [7, Chapter 2]. The complete proof can be found in Section A.3.

The distribution-free bound of Corollary 3 is obtained not directly as a corollary of Proposition 1, but as a consequence of its proof. (It is not enough to optimize the bound of Proposition 1 over the Δ_i, for it would yield an additional multiplicative factor of K.)

Corollary 3 (Distribution-free; Unif and EBA). *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round $K\lfloor n/K \rfloor$) ensures that*

$$\sup_{\nu_1,\dots,\nu_K} \mathbb{E}r_n \leqslant 2\sqrt{\frac{2K\ln K}{n}} \ ,$$

where the supremum is over all K-tuples $(\nu_1, \dots, \nu_K)$ of distributions over $[0,1]$.

PROOF. We extract from the proof of Proposition 1 that

$$\mathbb{P}\{J_n = i\} \leqslant \exp\left(-\frac{1}{2}\left\lfloor \frac{n}{K} \right\rfloor \Delta_i^2\right) \ ;$$

we now distinguish whether a given Δ_i is more or less than a threshold ε, use that $\sum \mathbb{P}\{J_n = i\} = 1$ and $\Delta_i \leqslant 1$ for all i, to write

$$\begin{aligned}\mathbb{E}r_n = \sum_{i=1}^{K} \Delta_i \, \mathbb{P}\{J_n = i\} &\leqslant \varepsilon + \sum_{i:\Delta_i>\varepsilon} \Delta_i \, \mathbb{P}\{J_n = i\} \qquad (2) \\ &\leqslant \varepsilon + \sum_{i:\Delta_i>\varepsilon} \Delta_i \, \exp\left(-\frac{1}{2}\left\lfloor \frac{n}{K} \right\rfloor \Delta_i^2\right) \\ &\leqslant \varepsilon + (K-1)\varepsilon \, \exp\left(-\frac{1}{2}\varepsilon^2 \left\lfloor \frac{n}{K} \right\rfloor\right) \ ,\end{aligned}$$

where the last inequality comes by function study, provided that $\varepsilon \geqslant 1/\lfloor n/K \rfloor$: for $C > 0$, the function $x \in [0,1] \mapsto x \exp(-Cx^2/2)$ is decreasing on $[1/\sqrt{C}, 1]$. Substituting $\varepsilon = \sqrt{(2\ln K)/\lfloor n/K \rfloor}$ concludes the proof.

4.2. Analysis of UCB(α) as an allocation strategy

We start by studying the recommendation given by the most played arm. A (distribution-dependent) bound is stated in Theorem 2; the bound does not involve any quantity depending on the Δ_i, but it only holds for rounds n large enough, a statement that does involve the Δ_i. Its interest is first that it is simple to read, and second, that the techniques used to prove it imply easily a second (distribution-free) bound, stated in Theorem 3 and which is comparable to Corollary 3.

Theorem 2 (Distribution-dependent; UCB(α) and MPA). *For $\alpha > 1$, the allocation strategy given by UCB(α) associated to the recommendation given by the most played arm ensures that*

$$\mathbb{E}r_n \leqslant \frac{K}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that $n \geqslant K + \dfrac{4K\alpha \ln n}{\Delta^2}$ and $n \geqslant K(K+2)$.

The polynomial rate in the upper bound above is not a coincidence according to the lower bound exhibited in Corollary 2. Here, surprisingly enough, this polynomial rate of decrease is distribution-free (but in compensation, the bound is only valid after a distribution-dependent time). This rate illustrates Theorem 1: the larger α, the larger the (theoretical bound on the) cumulative regret of UCB(α) but the smaller the simple regret of UCB(α) associated to the recommendation given by the most played arm.

Theorem 3 (Distribution-free; UCB(α) and MPA). *For $\alpha > 1$, the allocation strategy given by UCB(α) associated to the recommendation given by the most played arm ensures that, for all $n \geqslant K(K+2)$,*

$$\sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leqslant \sqrt{\frac{4K\alpha \ln n}{n-K}} + \frac{K}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)} = O\left(\sqrt{\frac{K\alpha \ln n}{n}}\right) ,$$

where the supremum is over all K-tuples $(\nu_1, \ldots, \nu_K)$ of distributions over $[0,1]$.

4.2.1. Proofs of Theorems 2 and 3

We start by a technical lemma from which the two theorems will follow easily.

Lemma 1. *Let $a_1, \ldots, a_K$ be real numbers such that $a_1 + \ldots + a_K = 1$ and $a_i \geqslant 0$ for all i, with the additional property that for all suboptimal arms i and all optimal arms i^*, one has $a_i \leqslant a_{i^*}$. Then for $\alpha > 1$, the allocation strategy*

given by UCB(α) associated to the recommendation given by the most played arm ensures that

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha-1}\sum_{i\neq i^*}(a_i n-1)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that, for all suboptimal arms i,

$$a_i n \geqslant 1+\frac{4\alpha\ln n}{\Delta_i^2} \quad \textit{and} \quad a_i n \geqslant K+2\ .$$

Proof. We first prove that whenever the most played arm J_n is different from an optimal arm i^*, then at least one of the suboptimal arms i is such that $T_i(n)\geqslant a_i n$. To do so, we prove the converse and assume that $T_i(n)<a_i n$ for all suboptimal arms. Then,

$$\left(\sum_{i=1}^K a_i\right)n = n = \sum_{i=1}^K T_i(n) < \sum_{i^*}T_{i^*}(n)+\sum_i a_i n$$

where, in the inequality, the first summation is over the optimal arms, the second one, over the suboptimal ones. Therefore, we get

$$\sum_{i^*} a_{i^*} n < \sum_{i^*} T_{i^*}(n)$$

and there exists at least one optimal arm i^* such that $T_{i^*}(n)>a_{i^*}n$. Since by definition of the vector $(a_1,\dots,a_K)$, one has $a_i\leqslant a_{i^*}$ for all suboptimal arms, it comes that $T_i(n)<a_i n\leqslant a_{i^*}n<T_{i^*}(n)$ for all suboptimal arms, and the most played arm J_n is thus an optimal arm.

Thus, using that $\Delta_i\leqslant 1$ for all i,

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} \leqslant \sum_{i:\Delta_i>0}\mathbb{P}\big\{T_i(n)\geqslant a_i n\big\}\ .$$

A side-result extracted from the proof of [1, proof of Theorem 7], see also [2, proof of Theorem 1], states that for all suboptimal arms i and all rounds $t\geqslant K+1$,

$$\mathbb{P}\Big\{I_t=i \ \text{ and } \ T_i(t-1)\geqslant \ell\Big\}\leqslant 2\,t^{1-2\alpha} \qquad \text{whenever} \qquad \ell\geqslant\frac{4\alpha\ln n}{\Delta_i^2}\ . \quad (3)$$

We denote by $\lceil x\rceil$ the upper integer part of a real number x. For a suboptimal arm i and since by the assumptions on n and the a_i, the choice $\ell=\lceil a_i n\rceil-1$ satisfies $\ell\geqslant K+1$ and $\ell\geqslant(4\alpha\ln n)/\Delta_i^2$,

$$\begin{aligned}\mathbb{P}\big\{T_i(n)\geqslant a_i n\big\} &= \mathbb{P}\big\{T_i(n)\geqslant\lceil a_i n\rceil\big\}\\ &\leqslant \sum_{t=\lceil a_i n\rceil}^n \mathbb{P}\Big\{T_i(t-1)=\lceil a_i n\rceil-1 \ \text{ and } \ I_t=i\Big\}\\ &\leqslant \sum_{t=\lceil a_i n\rceil}^n 2\,t^{1-2\alpha}\leqslant 2\int_{\lceil a_i n\rceil-1}^\infty v^{1-2\alpha}\,\mathrm{d}v\leqslant\frac{1}{\alpha-1}(a_i n-1)^{2(1-\alpha)}\ , \quad (4)\end{aligned}$$

where we used a union bound for the second inequality and (3) for the third inequality. A summation over all suboptimal arms i concludes the proof.

Proof (of Theorem 2). It consists in applying Lemma 1 with the uniform choice $a_i = 1/K$ and recalling that Δ is the minimum of the $\Delta_i > 0$.

Proof (of Theorem 3). We start the proof by using that $\sum \mathbb{P}\{J_n = i\} = 1$ and $\Delta_i \leqslant 1$ for all i, and can thus write

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} = \sum_{i=1}^{K} \Delta_i \, \mathbb{P}\{J_n = i\} \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \Delta_i \, \mathbb{P}\{J_n = i\} \ .$$

Since $J_n = i$ only if $T_i(n) \geqslant n/K$, we get

$$\mathbb{E}r_n \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \Delta_i \, \mathbb{P}\Big\{T_i(n) \geqslant \frac{n}{K}\Big\} \ .$$

Applying (4) with $a_i = 1/K$ leads to

$$\mathbb{E}r_n \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \frac{\Delta_i}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)} \ ,$$

where ε is chosen such that for all $\Delta_i > \varepsilon$, the condition

$$\ell \geqslant n/K - 1 \geqslant (4\alpha \ln n)/\Delta_i^2$$

is satisfied ($n/K - 1 \geqslant K + 1$ being satisfied by the assumption on n and K). The conclusion thus follows from taking, for instance,

$$\varepsilon = \sqrt{(4\alpha K \ln n)/(n - K)}$$

and upper bounding all remaining Δ_i by 1.

4.2.2. Other recommendation strategies

We discuss here the combination of UCB(α) with the two other recommendation strategies, namely, the choice of the empirical best arm and the use of the empirical distribution of plays.

Remark 2 (UCB(α) and EDP). We indicate in this remark from which results the corresponding bounds of Table 1 follow. As noticed in the beginning of Section 3, in the case of a recommendation formed by the empirical distribution of plays, the simple regret is bounded in terms of the cumulative regret as $\mathbb{E}r_n \leqslant \mathbb{E}R_n/n$. Now, [2, 1] show that the cumulative regret of UCB(α) is less than something of the form

$$\bigcirc \, \alpha \ln n + \frac{3K}{2} + \frac{K}{2(\alpha - 1)} \ ,$$

where $\bigcirc$ denotes a constant dependent on $\nu_1, \ldots, \nu_K$. The distribution-free bound on $\mathbb{E}R_n$ (and thus on $\mathbb{E}r_n$) follows from the control, yielded by (3) and a summation,

$$\mathbb{E}T_i(n) \leqslant \frac{4\alpha \ln n}{\Delta_i^2} + \frac{3}{2} + \frac{1}{2(\alpha-1)} \; ,$$

together with the concavity argument

$$\mathbb{E}R_n = \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{E}T_i(n) = \sum_{i:\Delta_i>0} \left(\Delta_i \sqrt{\mathbb{E}T_i(n)} \right) \sqrt{\mathbb{E}T_i(n)}$$

$$\leqslant \sqrt{4\alpha \ln n + \frac{3}{2} + \frac{1}{2(\alpha-1)}} \sum_{i:\Delta_i>0} \sqrt{\mathbb{E}T_i(n)} \leqslant \sqrt{\left(4\alpha \ln n + \frac{3}{2} + \frac{1}{2(\alpha-1)} \right) Kn} \, ,$$

where Jensen's inequality guaranteed that $\sum \sqrt{\mathbb{E}T_i(n)} \leqslant \sqrt{Kn}$.

Remark 3 (UCB(α) and EBA). We can rephrase the results of [12] as using UCB1 as an allocation strategy and forming a recommendation according to the empirical best arm. In particular, [12, Theorem 5] provides a distribution-dependent bound on the probability of not picking the best arm with this procedure and can be used to derive the following bound on the simple regret of UCB(α) combined with EBA: for all $n \geqslant 1$,

$$\mathbb{E}r_n \leqslant \sum_{i:\Delta_i>0} \frac{4}{\Delta_i} \left(\frac{1}{n} \right)^{\rho_\alpha \Delta_i^2/2}$$

where ρ_α is a positive constant depending on α only. The leading constants $1/\Delta_i$ and the distribution-dependent exponent make it not as useful as the one presented in Theorem 2. The best distribution-free bound we could get from this bound was of the order of $1/\sqrt{\rho_\alpha \ln n}$, to be compared to the asymptotic optimal $1/\sqrt{n}$ rate stated in Theorem 3.

5. Conclusions: Comparison of the bounds, simulation study

We first explain why, in some cases, the bound provided by our theoretical analysis in Lemma 1 (for UCB(α) and MPA) is better than the bound stated in Proposition 1 (for Unif and EBA). The central point in the argument is that the bound of Lemma 1 is of the form $\bigcirc n^{2(1-\alpha)}$, for some distribution-dependent constant $\bigcirc$, that is, it has a distribution-free convergence rate. In comparison, the bound of Proposition 1 involves the gaps Δ_i in the rate of convergence. Some care is needed in the comparison, since the bound for UCB(α) holds only for n large enough, but it is easy to find situations where for moderate values of n, the bound exhibited for the sampling with UCB(α) is better than the one for the uniform allocation. These situations typically involve a rather large number K of arms; in the latter case, the uniform allocation strategy only samples $\lfloor n/K \rfloor$ times each arm, whereas the UCB strategy focuses rapidly its

exploration on the best arms. A general argument is proposed in Section A.4 as well as a numerical example, showing that for moderate values of n, the bounds associated to the sampling with UCB(α) are better than the ones associated to the uniform sampling. This is further illustrated numerically, in the right part of Figure 4).

To make short the longer story described in this paper, one can distinguish three regimes, according to the value of the number of rounds n. The statements of these regimes (the ranges of their corresponding n) involve distribution-dependent quantifications, to determine which n are considered small, moderate, or large.

- For large values of n, uniform exploration is better (as shown by a combination of the lower bound of Corollary 2 and of the upper bound of Proposition 1).
- For moderate values of n, sampling with UCB(α) is preferable, as discussed just above (and in Section A.4).
- For small values of n, little can be said and the best bounds to consider are perhaps the distribution-free bounds, which are of the same order of magnitude for the two pairs of strategies.

We propose two simple experiments to illustrate our theoretical analysis; each of them was run on 10^4 instances of the problem and we plotted the average simple regret. This is an instance of the Monte-Carlo method and provides accurate estimators of the expected simple regret $\mathbb{E}r_n$.

The first experiment (upper plot of Figure 4) shows that for small values of n (here, $n \leqslant 80$), the uniform allocation strategy can have an interesting behavior. Of course the range of these "small" values of n can be made arbitrarily large by decreasing the gap Δ. The second one (lower plot of Figure 4) corresponds to the numerical example to be described in Section A.4. In both cases, the unclear picture for small values of n become clearer for moderate values and shows an advantage in favor of UCB–based strategies.

Remark 4. We mostly illustrated here the small and moderate n regimes. This is because for large n, the simple regret is usually very small, even below computer precision. Therefore, because of the chosen ranges, we do not see yet the uniform allocation strategy getting better than UCB–based strategies, a fact that is true however for large enough n. This has an important impact on the interpretation of the lower bound of Theorem 1. While its statement is in finite time, it should be interpreted as providing an asymptotic result only.

6. Pure exploration for $\mathcal{X}$–armed bandits (i.e., in topological spaces)

This section is of theoretical interest. We consider the $\mathcal{X}$-armed bandit problem, of, e.g., [5, 11], and (re)define the notions of cumulative and simple regret

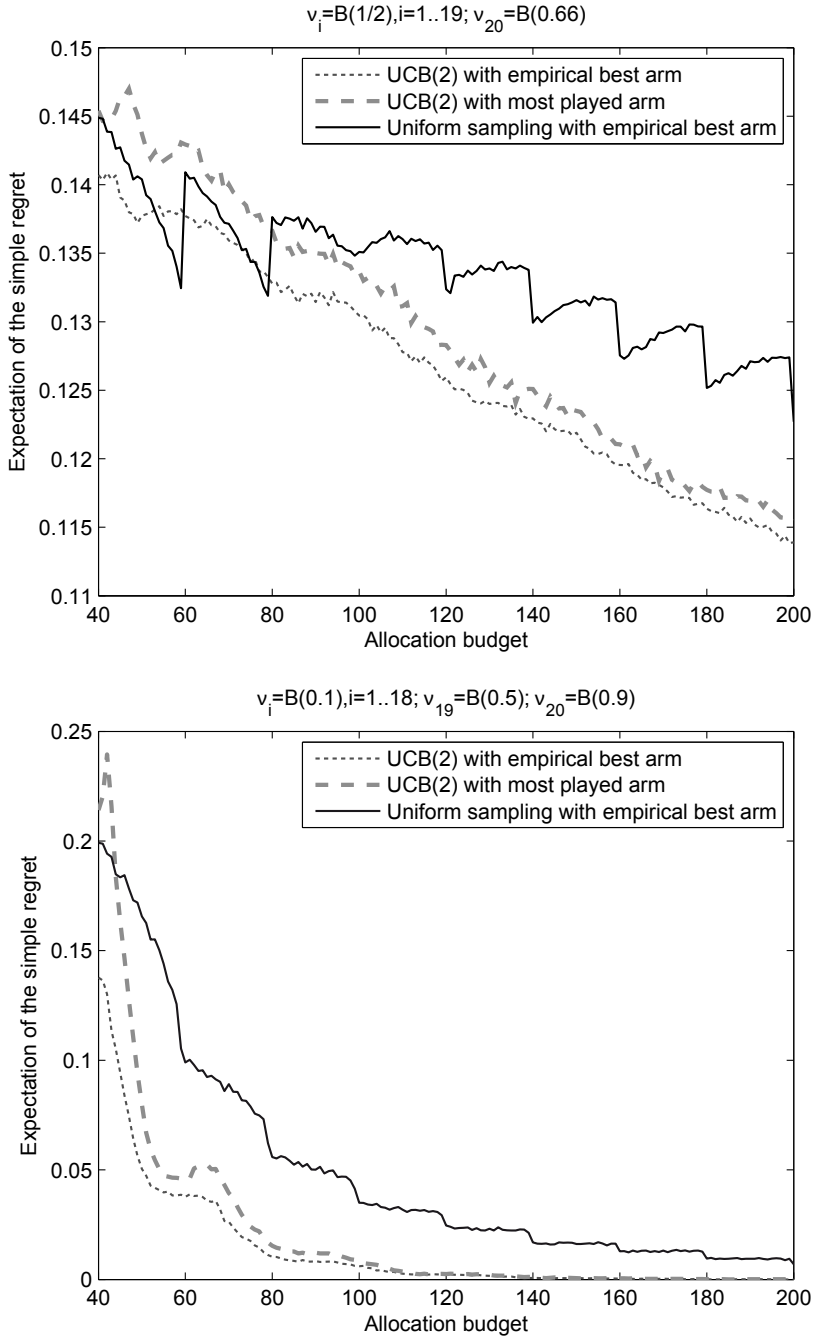

Figure 4: $K = 20$ arms with Bernoulli distributions of parameters indicated on top of each graph. x-axis: number of rounds n; y-axis: simple regrets $\mathbb{E}r_n$ (estimated by a Monte-Carlo method).

in this setting. We show that the cumulative regret can be minimized if and only if the simple regret can be minimized, and use this equivalence to characterize the metric spaces $\mathcal{X}$ in which the cumulative regret can be minimized: the separable ones. Here, in addition to its natural interpretation, the simple regret thus appears as a tool for proving results on the cumulative regret.

6.1. Description of the model of $\mathcal{X}$–armed bandits

We consider a bounded interval of $\mathbb{R}$, say $[0,1]$ again. We denote by $\mathcal{P}([0,1])$ the set of probability distributions over $[0,1]$. Similarly, given a topological space $\mathcal{X}$, we denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$. We then call environment on $\mathcal{X}$ any mapping $E : \mathcal{X} \to \mathcal{P}([0,1])$. We say that E is continuous if the mapping that associates to each $x \in \mathcal{X}$ the expectation $\mu(x)$ of $E(x)$ is continuous.

The $\mathcal{X}$–armed bandit problem is described in Figures 5 and 6. There, an environment E on $\mathcal{X}$ is fixed and we want various notions of regret to be small, given this environment.

Parameters: an environment $E : \mathcal{X} \to \mathcal{P}([0,1])$

For each round $t = 1, 2, \ldots,$

(1) the forecaster chooses a distribution $\varphi_t \in \mathcal{P}(\mathcal{X})$ and pulls an arm I_t at random according to φ_t;

(2) the environment draws the reward Y_t for that action, according to $E(I_t)$.

Goal:
Find an allocation strategy (φ_t) such that the cumulative regret

$$R_n = n \sup_{x \in \mathcal{X}} \mu(x) - \sum_{t=1}^{n} \mu(I_t)$$

is small (i.e., $o(n)$, in expectation).

Figure 5: The classical $\mathcal{X}$–armed bandit problem.

We consider now families of environments and say that a family $\mathcal{F}$ of environments is explorable–exploitable (respectively, explorable) if there exists a forecaster such that for any environment $E \in \mathcal{F}$, the expected cumulative regret $\mathbb{E}R_n$ (expectation taken with respect to E and all auxiliary randomizations) is $o(n)$ (respectively, $\mathbb{E}r_n = o(1)$). Of course, explorability of $\mathcal{F}$ is a milder requirement than explorability–exploitability of $\mathcal{F}$, as can be seen by considering the recommendation given by the empirical distribution of plays of Figure 3 and applying the same argument as the one used at the beginning of Section 3.

Parameters: an environment $E : \mathcal{X} \to \mathcal{P}([0,1])$

For each round $t = 1, 2, \ldots,$

(1) the forecaster chooses a distribution $\varphi_t \in \mathcal{P}(\mathcal{X})$ and pulls an arm I_t at random according to φ_t;

(2) the environment draws the reward Y_t for that action, according to $E(I_t)$;

(3) the forecaster outputs a recommendation $\psi_t \in \mathcal{P}(\mathcal{X})$;

(4) if the environment sends a stopping signal, then the game takes an end; otherwise, the next round starts.

Goal:
Find an allocation strategy (φ_t) and a recommendation strategy (ψ_t) such that the simple regret

$$r_n = \sup_{x \in \mathcal{X}} \mu(x) - \int_{\mathcal{X}} \mu(x) \, \mathrm{d}\psi_n(x)$$

is small (i.e., $o(1)$, in expectation).

Figure 6: The pure exploration problem for $\mathcal{X}$–armed bandits.

In fact, it can be seen that the two notions are equivalent, and this is why we will henceforth concentrate on explorability only, for which characterizations as the ones of Theorem 4 are simpler to exhibit and prove.

Lemma 2. *A family of environments $\mathcal{F}$ is explorable if and only if it is explorable–exploitable.*

The proof can be found in Section A.1. It relies essentially on designing a strategy suited for cumulative regret from a strategy minimizing the simple regret; to do so, exploration and exploitation occur at fixed rounds in two distinct phases and only the payoffs obtained during exploration rounds are fed into the base allocation strategy.

6.2. A positive result for metric spaces

We denote by $\mathcal{P}([0,1])^{\mathcal{X}}$ the family of all possible environments E on $\mathcal{X}$, and by $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ the subset of $\mathcal{P}([0,1])^{\mathcal{X}}$ formed by the continuous environments.

Example 1. Previous sections were about the family $\mathcal{P}([0,1])^{\mathcal{X}}$ of all environments over $\mathcal{X} = \{1, \ldots, K\}$ being explorable.

The main result concerning $\mathcal{X}$–armed bandit problems is formed by the following equivalences in metric spaces. It generalizes the result of Example 1.

Theorem 4. *Let $\mathcal{X}$ be a metric space. Then $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ is explorable if and only if $\mathcal{X}$ is separable.*

Corollary 4. *Let $\mathcal{X}$ be a set. $\mathcal{P}([0,1])^{\mathcal{X}}$ is explorable if and only if $\mathcal{X}$ is countable.*

The proofs can be found in Section A.2. Their main technical ingredient is that there exists a probability distribution over a metric space $\mathcal{X}$ giving a positive probability mass to all open sets if and only if $\mathcal{X}$ is separable. Then, whenever it exists, it allows some uniform exploration.

Acknowledgements

The authors acknowledge support by the French National Research Agency (ANR) under grants 08-COSI-004 "Exploration–exploitation for efficient resource allocation" (EXPLO/RA) and JCJC06-137444 "From applications to theory in learning and adaptive statistics" (ATLAS), as well as by the PASCAL Network of Excellence under EC grant no. 506778.

An extended abstract of the present paper appeared in the *Proceedings of the 20th International Conference on Algorithmic Learning Theory* (ALT'09).

A. Appendix

A.1. Proof of Lemma 2

Proof. In view of the comments before the statement of Lemma 2, we need only to prove that an explorable family $\mathcal{F}$ is also explorable–exploitable. We consider a pair of allocation (φ_t) and recommendation (ψ_t) strategies such that for all environments $E \in \mathcal{F}$, the simple regret satisfy $\mathbb{E}r_n = o(1)$, and provide a new strategy (φ'_t) such that its cumulative regret satisfies $\mathbb{E}R'_n = o(n)$ for all environments $E \in \mathcal{F}$.

It is defined informally as follows. At round $t = 1$, it uses $\varphi'_1 = \varphi_1$ and gets a reward Y_1. Based on this reward, the recommendation $\psi_1(Y_1)$ is formed and at round $t = 2$, the new strategy plays $\varphi'_2(Y_1) = \psi_1(Y_1)$. It gets a reward Y_2 but does not take it into account. It bases its choice $\varphi'_3(Y_1, Y_2) = \varphi_2(Y_1)$ only on Y_1 and gets a reward Y_3. Based on Y_1 and Y_3, the recommendation $\psi_2(Y_1, Y_3)$ is formed and played at rounds $t = 4$ and $t = 5$, i.e.,

$$\varphi'_4(Y_1, Y_2, Y_3) = \varphi'_5(Y_1, Y_2, Y_3, Y_4) = \psi_2(Y_1, Y_3) \ .$$

And so on: the sequence of distributions chosen by the new strategy is formed using the applications

$$\begin{array}{ll} \varphi_1, & \psi_1, \\ \varphi_2, & \psi_2, \ \psi_2, \\ \varphi_3, & \psi_3, \ \psi_3, \ \psi_3, \\ \varphi_4, & \psi_4, \ \psi_4, \ \psi_4, \ \psi_4, \\ \varphi_5, & \psi_5, \ \psi_5, \ \psi_5, \ \psi_5, \ \psi_5, \\ \dots & \end{array}$$

Formally, we consider regimes indexed by integers $t \geqslant 1$ and of length $1 + t$. The t–th regime starts at round

$$1 + \sum_{s=1}^{t-1}(1 + s) = t + \frac{t(t-1)}{2} = \frac{t(t+1)}{2} \ .$$

During this regime, the following distributions are used,

$$\varphi'_{t(t+1)/2+k} = \begin{cases} \varphi_t\Big(\big(Y_{s(s+1)/2}\big)_{s=1,\dots,t-1}\Big) & \text{if } k = 0; \\ \psi_t\Big(\big(Y_{s(s+1)/2}\big)_{s=1,\dots,t-1}\Big) & \text{if } 1 \leqslant k \leqslant t. \end{cases}$$

Note that we only keep track of the payoffs obtained when $k = 0$ in a regime.

The regret R'_n at round n of this strategy is as follows. We decompose n in a unique manner as

$$n = \frac{t(n)\big(t(n)+1\big)}{2} + k(n) \qquad \text{where} \qquad k(n) \in \big\{0, \dots, t(n)\big\} \ . \tag{5}$$

Then (using also the tower rule),

$$\mathbb{E}R'_n \leqslant t(n) + \Big(\mathbb{E}r_1 + 2\,\mathbb{E}r_2 + \ldots + \big(t(n)-1\big)\,\mathbb{E}r_{t(n)-1} + k(n)\,\mathbb{E}r_{t(n)}\Big)$$

where the first term comes from the time rounds when the new strategy used the base allocation strategy to explore and where the other terms come from the ones when it exploited. This inequality can be rewritten as

$$\frac{\mathbb{E}R'_n}{n} \leqslant \frac{t(n)}{n} + \frac{k(n)\,\mathbb{E}r_{t(n)} + \sum_{s=1}^{t(n)-1} s\,\mathbb{E}r_s}{n} \;,$$

which shows that $\mathbb{E}R'_n = o(n)$ whenever $\mathbb{E}r_s = o(1)$ as $s \to \infty$, since the first term in the right-hand side is of the order of $1/\sqrt{n}$ and the second one is a Cesaro average. This concludes that the exhibited strategy has a small cumulative regret for all environments of the family, which is thus explorable–exploitable.

A.2. Proof of Theorem 4 and its corollary

The key ingredient is the following characterization of separability (which relies on an application of Zorn's lemma); see, e.g., [4, Appendix I, page 216].

Lemma 3. *Let $\mathcal{X}$ be a metric space, with distance denoted by d. $\mathcal{X}$ is separable if and only if it contains no uncountable subset A such that*

$$\rho = \inf\big\{d(x,y) : x, y \in A\big\} > 0 \;.$$

Separability can then be characterized in terms of the existence of a probability distribution with full support. Though it seems natural, we did not see any reference to it in the literature and this is why we state it. (In the proof of Theorem 4, we will only use the straightforward direct part of the characterization.)

Lemma 4. *Let $\mathcal{X}$ be a metric space. There exists a probability distribution λ on $\mathcal{X}$ with $\lambda(V) > 0$ for all open sets V if and only if $\mathcal{X}$ is separable.*

Proof. We prove the converse implication first. If $\mathcal{X}$ is separable, we denote by $x_1, x_2, \ldots$ a dense sequence. If it is finite with length N, we let

$$\lambda = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$$

and otherwise,

$$\lambda = \sum_{i \geqslant 1} \frac{1}{2^i}\delta_{x_i} \;.$$

The result follows, since each open set V contains at least some x_i.

For the direct implication, we use Lemma 3 (and its notations). If $\mathcal{X}$ is not separable, then it contains uncountably many disjoint open balls, formed by the $B(a, \rho/2)$, for $a \in A$. If there existed a probability distribution λ with full support on $\mathcal{X}$, it would in particular give a positive probability to all these balls; but this is impossible, since there are uncountably many of them.

A.2.1. Separability of $\mathcal{X}$ implies explorability of the family $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$

The proof of the converse part of the characterization provided by Theorem 4 relies on a somewhat uniform exploration. We reach each open set of $\mathcal{X}$ in a geometric time.

PROOF. Since $\mathcal{X}$ is separable, there exists a probability distribution λ on $\mathcal{X}$ with $\lambda(V) > 0$ for all open sets V, as asserted by Lemma 4.

The proposed strategy is then constructed in a way similar to the one exhibited in Section A.4, in the sense that we also consider successives regimes, where the t–th of them has also length $1+t$. They use the following allocations,

$$\varphi_{t(t+1)/2+k} = \begin{cases} \lambda & \text{if } k = 0; \\ \delta_{I_{k(k+1)/2}} & \text{if } 1 \leqslant k \leqslant t. \end{cases}$$

Put in words, at the beginning of each regime, a new point $I_{t(t+1)/2}$ is drawn at random in $\mathcal{X}$ according to λ, and then, all previously drawn points $I_{s(s+1)/2}$, for $1 \leqslant s \leqslant t-1$, and the new point $I_{t(t+1)/2}$ are pulled again, one after the other.

The recommendations ψ_n are deterministic and put all probability mass on the best empirical arm among the first played $g(n)$ arms (where the function g will be determined by the analysis). Formally, for all $x \in \mathcal{X}$ such that

$$T_n(x) = \sum_{t=1}^{n} \mathbb{I}_{\{I_t = x\}} \geqslant 1 \ ,$$

one defines

$$\widehat{\mu}_n(x) = \frac{1}{T_n(x)} \sum_{t=1}^{n} Y_t \, \mathbb{I}_{\{I_t = x\}} \ .$$

Then,

$$\psi_n = \delta_{X_n^*} \quad \text{where} \quad X_n^* \in \operatorname*{argmax}_{1 \leqslant s \leqslant g(n)} \widehat{\mu}_n\big(I_{s(s+1)/2}\big)$$

(ties broken in some way, as usual; and $g(n)$ to be chosen small enough so that all considered arms have been played at least once). Note that exploration and exploitation appear in two distinct phases, as was the case already, for instance, in Section 4.1.

We now denote

$$\mu^* = \sup_{x \in \mathcal{X}} \mu(x) \quad \text{and} \quad \mu^*_{g(n)} = \max_{1 \leqslant s \leqslant g(n)} \mu\big(I_{s(s+1)/2}\big) \ ;$$

the simple regret can then be decomposed as

$$\mathbb{E}r_n = \mu^* - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big] = \Big(\mu^* - \mathbb{E}\Big[\mu^*_{g(n)}\Big]\Big) + \Big(\mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big]\Big) \ ,$$

where the first difference can be thought of as an approximation error, and the second one, as resulting from an estimation error. We now show that both differences vanish in the limit.

We first deal with the approximation error. We fix $\varepsilon > 0$. Since μ is continuous on $\mathcal{X}$, there exists an open set V such that

$$\forall x \in V, \qquad \mu^* - \mu(x) \leqslant \varepsilon \ .$$

It follows that

$$\begin{aligned}\mathbb{P}\Big\{\mu^* - \mu^*_{g(n)} > \varepsilon\Big\} &\leqslant \mathbb{P}\Big\{\forall\, s \in \big\{1, \dots, g(n)\big\}, \quad I_{s(s+1)/2} \notin V\Big\} \\ &\leqslant \big(1 - \lambda(V)\big)^{g(n)} \longrightarrow 0\end{aligned}$$

provided that $g(n) \to \infty$ (a condition that will be satisfied, see below). Since in addition, $\mu^*_{g(n)} \leqslant \mu^*$, we get

$$\limsup \ \mu^* - \mathbb{E}\Big[\mu^*_{g(n)}\Big] \leqslant \varepsilon \ .$$

For the difference resulting from the estimation error, we denote

$$I^*_n \in \operatorname*{argmax}_{1 \leqslant s \leqslant g(n)} \mu\big(I_{s(s+1)/2}\big)$$

(ties broken in some way). Fix an arbitrary $\varepsilon > 0$. We note that if for all $1 \leqslant s \leqslant g(n)$,

$$\Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| \leqslant \varepsilon \ ,$$

then (together with the definition of X^*_n)

$$\mu\big(X^*_n\big) \geqslant \widehat{\mu}_n\big(X^*_n\big) - \varepsilon \geqslant \widehat{\mu}_n\big(I^*_n\big) - \varepsilon \geqslant \mu\big(I^*_n\big) - 2\varepsilon \ .$$

Thus, we have proved the inequality

$$\mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 2\varepsilon + \mathbb{P}\Big\{\exists\, s \leqslant g(n), \Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon\Big\} \ . \tag{6}$$

We use a union bound and control each (conditional) probability

$$\mathbb{P}\Big\{\Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon \ \Big| \ \mathcal{A}_n\Big\} \tag{7}$$

for $1 \leqslant s \leqslant g(n)$, where $\mathcal{A}_n$ is the σ–algebra generated by the randomly drawn points $I_{k(k+1)/2}$, for those k with $k(k+1)/2 \leqslant n$. Conditionnally to them, $\widehat{\mu}_n\big(I_{s(s+1)/2}\big)$ is an average of a deterministic number of summands, which only depends on s, and thus, classical concentration-of-the-measure arguments can be used. For instance, the quantities (7) are bounded, via an application of Hoeffding's inequality (for i.i.d. random variables, see [10]), by

$$2\exp\Big(-2\,T_n\big(I_{s(s+1)/2}\big)\,\varepsilon^2\Big) \ .$$

We lower bound $T_n\big(I_{s(s+1)/2}\big)$. The point $I_{s(s+1)/2}$ was pulled twice in regime s, once in each regime $s+1,\dots,t(n)-1$, and maybe in $t(n)$, where n is decomposed again as in (5). That is,

$$T_n\big(I_{s(s+1)/2}\big) \geqslant t(n) - s + 1 \geqslant \sqrt{2n} - 1 - g(n) \ ,$$

since we only consider $s \leqslant g(n)$ and since (5) implies that

$$n \leqslant \frac{t(n)\,\big(t(n)+3\big)}{2} \leqslant \frac{\big(t(n)+2\big)^2}{2} \ , \quad \text{that is,} \ \ t(n) \geqslant \sqrt{2n} - 2 \ .$$

Substituting this in the Hoeffding's bound, integrating, and taking a union bound lead from (6) to

$$\mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 2\varepsilon + 2g(n)\,\exp\Big(-2\,\big(\sqrt{2n} - 1 - g(n)\big)\,\varepsilon^2\Big) \ .$$

Choosing for instance $g(n) = \sqrt{n}/2$ ensures that

$$\limsup\ \mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 2\varepsilon \ .$$

Summing up the two superior limits, we finally get

$$\limsup\ \mathbb{E}r_n \leqslant \limsup\ \mu^* - \mathbb{E}\Big[\mu^*_{g(n)}\Big] + \limsup\ \mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 3\varepsilon \ ;$$

since this is true for all arbitrary $\varepsilon > 0$, the proof is concluded.

A.2.2. Separability of $\mathcal{X}$ is a necessary condition

We now prove the direct part of the characterization provided by Theorem 4. It basically follows from the impossibility of a uniform exploration, as asserted by Lemma 4.

PROOF. Let $\mathcal{X}$ be a non-separable metric space (with distance denoted by d). Let A be an uncountable set and let $\rho > 0$ be defined as in Lemma 3; in particular, the balls $B(a, \rho/2)$ are disjoint, for $a \in A$.

We now consider the subset of $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ formed by the environments E_a defined as follows. They are indexed by $a \in A$ and their corresponding expectations are given by

$$\mu_a : x \in \mathcal{X} \longmapsto \left(1 - \frac{d(x,a)}{\rho/2}\right)^+ \ .$$

Note that μ_a is continuous, that $\mu_a(x) > 0$ for all $x \in B(a, \rho/2)$ but $\mu_a(x) = 0$ for all $x \in \mathcal{X}\backslash B(a, \rho/2)$; that the best arm is a and gets a reward $\mu^*_a = \mu_a(a) = 1$. The associated environment E_a is deterministic, in the sense that it is defined as $E_a(x) = \delta_{\mu_a(x)}$.

We fix a forecaster and denote by $\mathbb{E}_a$ the expectation under environment E_a with respect with the auxiliary randomizations used by the forecaster. By construction of μ_a,

$$\mathbb{E}_a r_n = 1 - \mathbb{E}_a\left[\int_{\mathcal{X}} \mu_a(x)\,\mathrm{d}\psi_n(x)\right] \geqslant 1 - \mathbb{E}_a\Big[\psi_n\big(B(a,\rho/2)\big)\Big]\ .$$

We now show the existence of a non-empty set A' such that for all $a \in A'$ and $n \geqslant 1$,

$$\mathbb{E}_a\Big[\psi_n\big(B(a,\rho/2)\big)\Big] = 0\ ; \tag{8}$$

this indicates that $\mathbb{E}_a r_n = 1$ for all $n \geqslant 1$ and $a \in A'$, thus preventing in particular $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ from being explorable by the fixed forecaster.

The set A' is constructed by studying the behavior of the forecaster under the environment E_0 yielding deterministic null rewards throughout the space, i.e., associated to the expectations $x \in \mathcal{X} \mapsto \mu_0(x) = 0$. In the first round, the forecaster chooses a deterministic distribution $\varphi_1 = \varphi_1^0$ over $\mathcal{X}$, picks I_1 at random according to φ_1^0, gets a deterministic payoff $Y_1 = 0$, and finally recommends $\psi_1^0(I_1) = \psi_1(I_1, Y_1)$ (which depends on I_1 only, since the obtained payoffs are all null). In the second round, it chooses an allocation $\psi_2^0(I_1)$ (that depends only on I_1, for the same reasons as before), picks I_2 at random according to $\psi_2^0(I_1)$, gets a null reward, and recommends $\psi_2^0(I_1, I_2)$; and so on.

We denote by $\mathbb{A}$ the probability distribution giving the auxiliary randomizations used to draw the I_t at random, and for all integers t and all measurable applications

$$\nu : (x_1, \ldots, x_t) \in \mathcal{X}^t \longmapsto \nu(x_1, \ldots, x_t) \in \mathcal{P}(\mathcal{X})$$

we introduce the distributions $\mathbb{A}\cdot\nu \in \mathcal{P}(\mathcal{X})$ defined as follows. For all measurable sets $V \subseteq \mathcal{X}$,

$$\mathbb{A}\cdot\nu(V) = \mathbb{E}_{\mathbb{A}}\left[\int_{\mathcal{X}} \mathbb{I}_V\,\mathrm{d}\nu(I_1, \ldots, I_t)\right]\ .$$

Now, let B_n and C_n be defined as the at most countable sets of a such that, respectively, $\mathbb{A}\cdot\varphi_n^0$ and $\mathbb{A}\cdot\psi_n^0$ give a positive probability mass to $B(a,\rho/2)$; we recall that the latter is the support of the expectation mapping μ_a. Then, let

$$A' = A \setminus \left(\bigcup_{n\geqslant 1} B_n \cup \bigcup_{n\geqslant 1} C_n\right)$$

be the uncountable, thus non empty, set of those elements of A which are in no B_n or C_n.

By construction, for all $a \in A'$, the forecaster then behaves similarly under the environments E_a and E_0, since it only gets null rewards (a is in no B_n); this similar behavior means formally that for all measurable sets $V \subseteq \mathcal{X}$ and all $n \geqslant 1$,

$$\mathbb{E}_a\big[\varphi_n(V)\big] = \mathbb{A}\cdot\varphi_n^0(V) \quad \text{and} \quad \mathbb{E}_a\big[\psi_n(V)\big] = \mathbb{A}\cdot\psi_n^0(V)\ .$$

In particular, since a is in no C_n, it hits in no recommendation ψ_n the ball $B(a,\rho/2)$, which is exactly what remained to be proved, see (8).

A.2.3. The countable case of Corollary 4

We adopt an "à la Bourbaki" approach and derive this special case from the general theory.

PROOF. We endow $\mathcal{X}$ with the discrete topology, i.e., choose the distance

$$d(x,y) = \mathbb{I}_{\{x \neq y\}} \ .$$

Then, all applications defined on $\mathcal{X}$ are continuous; in particular,

$$\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big) = \mathcal{P}([0,1])^{\mathcal{X}} \ .$$

In addition, $\mathcal{X}$ is then separable if and only if it is countable. The result thus follows immediately from Theorem 4.

A.2.4. An additional remark

In this paper, we mostly consider non-uniform bounds (bounds that are individual as far as the environments are concerned). As for uniform bounds, i.e., bounds on quantities of the form

$$\sup_{E \in \mathcal{F}} \mathbb{E}R_n \qquad \text{or} \qquad \sup_{E \in \mathcal{F}} \mathbb{E}r_n$$

for some family $\mathcal{F}$, two observations can be made.

First, it is easy to see that no sublinear uniform bound can be obtained for the family of all continuous environments, as soon as there exists infinitely many disjoint open balls.

However one can exhibit such sublinear uniform bounds in some specific scenarios; for instance, when $\mathcal{X}$ is totally bounded and $\mathcal{F}$ is formed by continuous functions with a common bounded Lipschitz constant.

A.3. Proof of the second statement of Proposition 1

We use below the notations introduced in the proof of the first statement of Proposition 1.

PROOF. Since some regret is suffered only when an arm with suboptimal expectation has the best empirical performance,

$$\mathbb{E}r_n \leqslant \left(\max_{i=1,\dots,K} \Delta_i \right) \mathbb{P}\left\{ \max_{i:\Delta_i > 0} \widehat{\mu}_{i,n} \geqslant \widehat{\mu}_{i^*,n} \right\} \ .$$

Now, the quantity of interest can be rewritten as

$$\left\lfloor \frac{n}{K} \right\rfloor \left(\max_{i:\Delta_i > 0} \widehat{\mu}_{i,n} - \widehat{\mu}_{i^*,n} \right) = f\Big(\vec{X}_1, \dots, \vec{X}_{\lfloor \frac{n}{K} \rfloor} \Big)$$

for some function f, where for all $s = 1, \dots, \lfloor n/K \rfloor$, we denote by $\vec{X}_s$ the vector $(X_{1,s}, \dots, X_{K,s})$. ($f$ is defined as a maximum of at most $K-1$ sums of differences.) We apply the method of bounded differences, see [16], see also

[7, Chapter 2]. It is straightforward that, since all random variables of interest take values in $[0,1]$, the bounded differences condition is satisfied with ranges all equal to 2. Therefore, the indicated concentration inequality states that

$$\mathbb{P}\left\{\left(\max_{i:\Delta_i>0}\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}\right)-\mathbb{E}\left[\max_{i:\Delta_i>0}\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}\right]\geqslant\varepsilon\right\}\leqslant\exp\left(-\frac{2\lfloor n/K\rfloor\varepsilon^2}{4}\right)$$

for all $\varepsilon>0$. We choose

$$\varepsilon=-\mathbb{E}\left[\max_{i:\Delta_i>0}\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}\right]\geqslant\min_{i:\Delta_i>0}\Delta_i-\mathbb{E}\left[\max_{i:\Delta_i>0}\left\{\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}+\Delta_i\right\}\right]$$

(where we used that the maximum of K first quantities plus the minimum of K other quantities is less than the maximum of the K sums). We now argue that

$$\mathbb{E}\left[\max_{i:\Delta_i>0}\left\{\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}+\Delta_i\right\}\right]\leqslant\sqrt{\frac{2\ln K}{\lfloor n/K\rfloor}}\;;$$

this is done by a classical argument, using bounds on the moment generating function of the random variables of interest. Consider

$$Z_i=\lfloor n/K\rfloor\left(\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}+\Delta_i\right)$$

for all $i=1,\ldots,K$. Independence and Hoeffding's lemma (see, e.g., [7, Chapter 2]) imply that for all $\lambda>0$,

$$\mathbb{E}\left[e^{\lambda Z_i}\right]\leqslant\exp\left(-\frac{1}{2}\lambda^2\lfloor n/K\rfloor\right)$$

(where we used again that Z_i is given by a sum of random variables bounded between -1 and 1). A well-known inequality for maxima of subgaussian random variables (see, again, [7, Chapter 2]) then yields

$$\mathbb{E}\left[\max_{i=1,\ldots,K}Z_i\right]\leqslant\sqrt{2\lfloor n/K\rfloor\ln K}\;,$$

which leads to the claimed upper bound. Putting things together, we get that for the choice

$$\varepsilon=-\mathbb{E}\left[\max_{i:\Delta_i>0}\widehat{\mu}_{i,n}-\widehat{\mu}_{i^*,n}\right]\geqslant\min_{i:\Delta_i>0}\Delta_i-\sqrt{\frac{2\ln K}{\lfloor n/K\rfloor}}>0$$

(for n sufficiently large, a statement made precise below), we have

$$\begin{aligned}\mathbb{P}\left\{\max_{i:\Delta_i>0}\widehat{\mu}_{i,n}\geqslant\widehat{\mu}_{i^*,n}\right\}&\leqslant\exp\left(-\frac{2\lfloor n/K\rfloor\varepsilon^2}{4}\right)\\&\leqslant\exp\left(-\frac{1}{2}\left\lfloor\frac{n}{K}\right\rfloor\left(\min_{i:\Delta_i>0}\Delta_i-\sqrt{\frac{2\ln K}{\lfloor n/K\rfloor}}\right)^2\right)\;.\end{aligned}$$

The result follows for n such that

$$\min_{i:\Delta_i>0} \Delta_i - \sqrt{\frac{2\ln K}{\lfloor n/K \rfloor}} \geqslant \frac{1}{2} \min_{i:\Delta_i>0} \Delta_i \ ;$$

the second part of the theorem indeed only considers such n.

A.4. Detailed discussion of the heuristic arguments presented in Section 5

We first state the following corollary to Lemma 1.

Theorem 5. *The allocation strategy given by UCB(α) (where $\alpha > 1$) associated to the recommendation given by the most played arm ensures that*

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha - 1} \sum_{i \neq i^*} \left(\frac{\beta n}{\Delta_i^2} - 1 \right)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that

$$\frac{n}{\ln n} \geqslant \frac{4\alpha + 1}{\beta} \quad \textit{and} \quad n \geqslant \frac{K+2}{\beta} (\Delta')^2 \ ,$$

where $\Delta' = \max_i \Delta_i$ and we denote by K^* the number of optimal arms and

$$\beta = \frac{1}{\dfrac{K^*}{\Delta^2} + \displaystyle\sum_{i \neq i^*} \frac{1}{\Delta_i^2}} \ .$$

Proof. We apply Lemma 1 with the choice $a_i = \beta/\Delta_i^2$ for all suboptimal arms i and $a_{i^*} = \beta/\Delta^2$ for all optimal arms i^*, where β denotes the renormalization constant.

For illustration, consider the case when there is one optimal arm, one Δ–suboptimal arm and $K-2$ arms that are 2Δ–suboptimal. Then

$$\frac{1}{\beta} = \frac{2}{\Delta^2} + \frac{K-2}{(2\Delta)^2} = \frac{6+K}{4\Delta^2} \ ,$$

and the previous bound of Theorem 5 implies that

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha - 1} \left(\frac{4n}{6+K} - 1 \right)^{2(1-\alpha)} + \frac{K-2}{\alpha - 1} \left(\frac{n}{6+K} - 1 \right)^{2(1-\alpha)} \tag{9}$$

for all n sufficiently large, e.g.,

$$n \geqslant \max \left\{ (K+2)(6+K),\ (4\alpha+1) \left(\frac{6+K}{4\Delta^2} \right) \ln n \right\} \ . \tag{10}$$

Now, the upper bound on $\mathbb{E}r_n$ given in Proposition 1 for the uniform allocation associated to the recommendation provided by the empirical best arm is larger than

$$\Delta e^{-\Delta^2 \lfloor n/K \rfloor/2} , \qquad \text{for all } n \geqslant K.$$

Thus for n moderately large, e.g., such that $n \geqslant K$ and

$$\lfloor n/K \rfloor \leqslant (4\alpha + 1) \left(\frac{6+K}{4\Delta^2} \right) \frac{\ln n}{K} , \tag{11}$$

the bound for the uniform allocation is at least

$$\Delta \exp \left(-\Delta^2 (4\alpha + 1) \left(\frac{6+K}{4\Delta^2} \right) \frac{\ln n}{2K} \right) = \Delta n^{-(4\alpha+1)(6+K)/8K} ,$$

which may be much worse than the upper bound (9) for the UCB(α) strategy whenever K is large, as can be seen by comparing the exponents $-2(\alpha - 1)$ versus $-(4\alpha + 1)(6 + K)/8K$.

To illustrate this numerically (though this is probably not the most convincing choice of the parameters), consider the case when $\Delta = 0.4$, $K = 20$, and $\alpha = 4$. Then $n = 6020$ satisfies (10) and (11). For these parameters, the upper bound (9) for the UCB(α) strategy is 4.00×10^{-14}, which is much smaller than the one for the uniform allocation, which is larger than 1.45×10^{-11}.

The reason is that the uniform allocation strategy only samples $\lfloor n/K \rfloor$ each arm, whereas the UCB strategy focuses rapidly its exploration on the better arms.

References

[1] J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

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

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[4] P. Billingsley. *Convergence of Probability Measures*. Wiley and Sons, 1968.

[5] S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvari. Online optimization in $\mathcal{X}$-armed bandits. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 201–208, 2009.

[6] P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 67–74, 2007.

[7] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[8] E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 255–270, 2002.

[9] S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in Monte-Carlo go. Technical Report RR-6062, INRIA, 2006.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[11] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems*, 2004.

[12] L. Kocsis and Cs. Szepesvari. Bandit based Monte-carlo planning. In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.

[13] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[14] O. Madani, D. Lizotte, and R. Greiner. The budgeted multi-armed bandit problem. In *Proceedings of the 17th Annual Conference on Computational Learning Theory*, pages 643–645, 2004. Open problems session.

[15] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

[16] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148 – 188, 1989. Cambridge University Press.

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

[18] K. Schlag. Eleven tests needed for a recommendation. Technical Report ECO2006/2, European University Institute, 2006.