Pure Exploration for Multi-Armed Bandit Problems

Sébastien Bubeck, Rémi Munos, Gilles Stoltz

Pure Exploration in Multi-Armed Bandits Problems

Sébastien Bubeck
SequeL Project, INRIA Lille, France
sebastien.bubeck@inria.fr

Rémi Munos
SequeL Project, INRIA Lille, France
remi.munos@inria.fr

Gilles Stoltz
Ecole Normale Supérieure, HEC Paris, CNRS, France
gilles.stoltz@ens.fr

Abstract

We consider the framework of stochastic multi-armed bandit problems and study the possibilities and limitations of strategies that explore sequentially the arms. The strategies are assessed in terms of their simple regrets, 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 simple and cumulative regrets. 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. We then refine this statement.

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 [Rob52], 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 strategy is given by its cumulative regret, the sum of differences between the expected reward of the best arm and the obtained rewards. Typical good strategies, like the UCB strategies of [ACBF02], 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 recommendation, formed by a probability distribution over the arms. 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 [GWMT06]). A given time, i.e., a given amount of CPU times is given to the player to explore the possible outcome of a 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 [KS06] and the BAST strategy of [CM07]. 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., [Kle04, BMSS09]. 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 re-

sources (time, budget, or the number of patients to be included) – that is, we seek for anytime performance. The problem of pure exploration presented above was referred to as "budgeted multi-armed bandit problem" in the open problem [MLG04]. [Sch06] solves the pure exploration problem in a minmax sense for the case of two arms only and rewards given by probability distributions over $[0,1]$. [EDMM02] and [MT04] 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 in the statistical literature is the identification of the best arm (with high probability). However, the binary assessment criterion used there (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.

Problem setup, notation

We consider a sequential decision problem for multi-armed bandits, where a forecaster plays against a stochastic environment. $K \geqslant 2$ arms, denoted by $j = 1, \ldots, K$, are available and the j–th of them is parameterized by a probability distribution ν_j over $[0,1]$ (with expectation μ_j); at those rounds when it is pulled, its associated reward is drawn at random according to ν_j, independently of all previous rewards. For each arm j and all time rounds $n \geqslant 1$, we denote by $T_j(n)$ the number of times j was pulled from rounds 1 to n, and by $X_{j,1}, X_{j,2}, \ldots, X_{j,T_j(n)}$ 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. He may resort to a randomized strategy, which, based on past rewards, prescribes a probability distribution $\varphi_t \in \mathcal{P}\{1, \ldots, K\}$ (where we denote by $\mathcal{P}\{1, \ldots, K\}$ the set of all probability distributions over the indexes of the arms). In that case, I_t is drawn at random according to the probability distribution φ_t and 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 (φ_t) is referred to as an allocation strategy. The primary task is to output at the end of each round t a recommendation $\psi_t \in \mathcal{P}\{1, \ldots, K\}$ to be used to form a randomized play in a one-shot instance if/when the environment sends some stopping signal meaning that the exploration phase is over. The sequence (ψ_t) is referred to as a recommendation strategy. Figure 1 summarizes the description of the sequential game and points out that the information available to the forecaster for choosing φ_t, respectively ψ_t, is formed by the $X_{j,s}$ for $j = 1, \ldots, K$ and $s = 1, \ldots, T_j(t-1)$, respectively, $s = 1, \ldots, T_j(t)$.

As we are only interested in the performances of the recommendation strategy (ψ_t), we call this problem the pure exploration problem for multi-armed bandits and evaluate the strategies through their simple regrets. The simple regret of a recommendation $\psi_t = (\psi_{j,t})_{j=1,\ldots,N}$ is defined as the expected regret on a one-shot instance of the game, if a random action is taken according to ψ_t. Formally,

$$r_t = r\big(\psi_t\big) = \mu^* - \mu_{\psi_t} \qquad \text{where} \quad \mu^* = \mu_{j^*} = \max_{j=1,\ldots,K} \mu_j$$
$$\text{and} \qquad \mu_{\psi_t} = \sum_{j=1,\ldots,K} \psi_{j,t}\, \mu_j$$

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

$$\Delta = \min_{j:\Delta_j > 0} \Delta_j \ .$$

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

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

A popular treatment of the multi-armed bandit problems is to construct forecasters ensuring that $\mathbb{E}R_n = o(n)$, see, e.g., [LR85] or [ACBF02], and even $R_n = o(n)$ a.s., as follows, e.g., from [ACBFS02, Theorem 6.3] together with a martingale argument. The quantities $r'_t = \mu^* - \mu_{I_t}$ are sometimes called instantaneous regrets. They differ from the simple regrets 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 simple and cumulative regrets. Intuitively, an efficient allocation strategy for the simple regret should rely on some exploration–exploitation trade-off. Our main contribution (Theorem 1, Section 2) 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 Sections 3 and 4, we then

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 $\varphi_t \in \mathcal{P}\{1, \ldots, K\}$ and pulls an arm I_t at random according to φ_t;

(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 $\psi_t \in \mathcal{P}\{1, \ldots, K\}$;

(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.

refine this statement and illustrate it by simulations. In particular, we show how, despite all, strategies designed for the cumulative regret can outperform (for moderate values of n) strategies with optimal rates of convergence for the simple regret. Finally, in Section 5, we consider the setting of arms indexed by a metric space and discuss a necessary and sufficient condition for the existence of forecasters with small simple or cumulative regrets.

2 The smaller the cumulative regret, the larger the simple regret

It is immediate that for the recommendation formed by the empirical distribution of plays of Figure 3,

$$\psi_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{I_t} \ ,$$

the regrets satisfy $r_n = R_n/n$; therefore, upper bounds on $\mathbb{E}R_n$ lead to upper bounds on $\mathbb{E}r_n$. We show here that upper bounds on $\mathbb{E}R_n$ also lead to lower bounds on $\mathbb{E}r_n$: the better the guaranteed upper bound on $\mathbb{E}R_n$, the worst the lower bound on $\mathbb{E}r_n$, no matter what the recommendation strategies ψ_n are.

This is interpreted as a variation of the "classical" tradeoff between exploration and exploitation. Here, while the recommendation strategies ψ_n rely only on the exploitation of the results of the preliminary exploration phase, the design of the allocation policies φ_n 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 amount of exploration needed. In particular, to have an 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 Bernoulli distributions $\nu_1, \ldots, \nu_K$ over the rewards.

Theorem 1 (Main result) *For all allocation strategies (φ_t) and all functions $\varepsilon : \{1, 2, \ldots\} \to \mathbb{R}$ such that*

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

the simple regret of all recommendation strategies (ψ_t) based on the allocation strategies (φ_t) is such that

> *for all sets of $K \geqslant 3$ (distinct, Bernoulli) distributions on the rewards, all different from a Dirac distribution at 1, there exists a constant $D \geqslant 0$ with*
>
> $$\mathbb{E}r_n \geqslant \frac{\Delta}{2}\, e^{-D\varepsilon(n)}$$
>
> *(up to a relabeling $\nu_1, \ldots, \nu_K$ of the considered distributions into $\nu_{\pi(1)}, \ldots, \nu_{\pi(K)}$ for some permutation π).*

Since the cumulative regrets are always bounded by n, one gets the following.

Corollary 2 *For allocation strategies (φ_t), all recommendation strategies (ψ_t), and all sets of $K \geqslant 3$ (distinct, Bernoulli) distributions on the rewards, there exist two constants $\beta > 0$ and $\gamma \geqslant 0$ such that, up to relabeling,*

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

To get further the point of Theorem 1, one should keep in mind that the typical (distribution-dependent) rate of growth of the cumulative regrets of good algorithms, e.g., UCB1 of [ACBF02], is $\varepsilon(n) = \ln n$. This, as asserted in [LR85], is the optimal rate. But the recommendation strategies based on such allocation strategies are bound to suffer a simple regret that decreases at best polynomially fast. We state this result for the slight modification UCB(p) of UCB1 stated in Figure 2; its proof relies on noting that it achieves a cumulative regret bounded by $\varepsilon(n) = p \ln n$.

Corollary 3 *The allocation strategy (φ_t) given by the forecaster UCB(p) of Figure 2 ensures that for all recommendation strategies (ψ_t) and all sets of $K \geqslant 3$ (distinct, Bernoulli) distributions on the rewards, there exist two constants $\beta > 0$ and $\gamma \geqslant 0$ such that, up to relabeling,*

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

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 allocation strategies (φ_t) and a corresponding function ε such that the assumption of the theorem is satisfied. We consider below a set of $K \geqslant 3$ (distinct) Bernoulli distributions; actually, we only use below that their parameters are (up to a first relabeling) such that $1 > \mu_1 > \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_K \geqslant 0$ and $\mu_2 > \mu_K$ (thus, $\mu_2 > 0$).

Another layer of notation is needed. It depends on a given permutation 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 K-tuple of distributions overs the arms $\nu_1, \ldots, \nu_K$. 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 similar notation up to a relabeling. $\mathbb{P}_\sigma$ 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 j–th best arm is located in the $\sigma(j)$–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 $\nu_{\sigma^{-1}(j)}$, 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 a proof in six steps.

Step 1 lower bounds by an average the maximum of the simple regrets obtained by relabeling,

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

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

$$C_{j,n} = \sum_{t=1}^{T_j(n)} X_{j,t}$$

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

$$\begin{aligned} &\mathbb{E}_{\sigma} \left[1 - \psi_{\sigma(1),n} \right] \\ &\quad \geqslant \mathbb{E}_{\sigma} \left[\left(1 - \psi_{\sigma(1),n} \right) \mathbb{I}_{\{C_{\sigma(1),n}=0\}} \right] \\ &\quad = \mathbb{E}_{\sigma} \Big[\left(1 - \psi_{\sigma(1),n} \right) \mid C_{\sigma(1),n} = 0 \Big] \times \mathbb{P}_{\sigma} \left\{ C_{\sigma(1),n} = 0 \right\} \\ &\quad = \mathbb{E}_{1,\sigma} \Big[\left(1 - \psi_{\sigma(1),n} \right) \Big] \, \mathbb{P}_{\sigma} \left\{ C_{\sigma(1),n} = 0 \right\} . \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}_{1,\sigma} \Big[\left(1 - \psi_{\sigma(1),n} \right) \Big] \\ &\quad \geqslant \mathbb{E}_{1,\sigma} \Big[\left(1 - \psi_{\sigma(1),n} \right) \mid C_{\sigma(K),n} = 0 \Big] \\ &\qquad \times \mathbb{P}_{1,\sigma} \left\{ C_{\sigma(K),n} = 0 \right\} \\ &\quad = \mathbb{E}_{K,\sigma} \Big[\left(1 - \psi_{\sigma(1),n} \right) \Big] \, \mathbb{P}_{1,\sigma} \left\{ C_{\sigma(K),n} = 0 \right\} \end{aligned}$$

and therefore,

$$\begin{aligned} &\mathbb{E}_{\sigma} \left[1 - \psi_{\sigma(1),n} \right] \qquad (1) \\ &\quad \geqslant \mathbb{E}_{K,\sigma} \big[\left(1 - \psi_{\sigma(1),n} \right) \big] \times \mathbb{P}_{1,\sigma} \left\{ C_{\sigma(K),n} = 0 \right\} \\ &\qquad \times \mathbb{P}_{\sigma} \left\{ C_{\sigma(1),n} = 0 \right\} . \end{aligned}$$

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

$$\begin{aligned} &\mathbb{P}_{1,\sigma} \left\{ C_{\sigma(K),n} = 0 \right\} \\ &= \mathbb{E}_{1,\sigma} \Big[\left(1 - \mu_K \right)^{T_{\sigma(K)}(n)} \Big] \geqslant \left(1 - \mu_K \right)^{\mathbb{E}_{1,\sigma} T_{\sigma(K)}(n)} , \end{aligned}$$

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 sub-optimal arm $\sigma(K)$ is pulled under $\mathbb{P}_{1,\sigma}$ 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

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

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

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

We denote by $W_n = (I_1, Y_1, \ldots, I_n, Y_n)$ the history of actions pulled and obtained payoffs up to time n. What follows is reminiscent of the techniques used in [MT04]. 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,

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

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,

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

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,

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

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,

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

■

3 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, strategies not pulling each arm a linear number of the times in the exploration phase can have interesting simple regrets. To do so, and because of space constraints, we consider only two allocation strategies (the uniform allocation and the variant UCB(p) of UCB1 where the quantile factor may be a parameter) and 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 Figures 2 and 3.

Parameters: K arms

Uniform allocation
Plays all arms one after the other
For each round $t = 1, 2, \ldots,$

use $\varphi_t = \delta_{[t \bmod K]}$, where $[t \bmod K]$ denotes the value of t modulo K.

UCB(p)
First plays each arm once and then the one with the best upper confidence bound
Parameter: quantile factor p
For rounds $t = 1, \ldots, K$, play $\varphi_t = \delta_t$
For each round $t = K+1, K+2, \ldots,$

(1) compute, for all $j = 1, \ldots, K$, the quantities

$$\widehat{\mu}_{j,t-1} = \frac{1}{T_j(t-1)} \ \sum_{s=1}^{T_j(t-1)} X_{j,s} \ ;$$

(2) use $\varphi_t = \delta_{j^*_{t-1}}$, where

$$j^*_{t-1} \in \operatorname*{argmax}_{j=1,\ldots,K} \widehat{\mu}_{j,t-1} + \sqrt{\frac{p \ln(t-1)}{T_j(t-1)}}$$

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

Figure 2: Two allocation strategies.

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 $X_{j,1}, \ldots, X_{j,T_j(n)}$, for $j = 1, \ldots, n$

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

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

and forms a deterministic recommendation (conditionally to the history),

$$\psi_n = \delta_{J^*_n} \qquad \text{where} \quad J^*_n \in \operatorname*{argmax}_j \widehat{\mu}_{j,n}$$

(ties broken in some way).

Most played arm (MPA)
Forms a deterministic recommendation (conditionally to the history),

$$\psi_n = \delta_{J^*_n} \qquad \text{where} \quad J^*_n \in \operatorname*{argmax}_{j=1,\ldots,N} T_j(n) \ .$$

(ties broken in some way).

Empirical distribution of plays (EDP)
Draws a recommendation using the probability distribution

$$\psi_n = \frac{1}{n} \sum_{t=1}^{n} \delta_{I_t} \ .$$

Figure 3: Three recommendation strategies.

Before diving into the technical statements, we point out some inherent issues in exhibiting upper bounds for the simple regret. A first delicate case occurs when there is only one optimal arm and the gaps for all others take the common values Δ, for some small Δ. It seems rather intuitive that one can then not do anything better than assigning the same number of pulls to all the arms; the formalization of this intuition leads to the distribution-free lower bound of [ACBFS02]. Thus, for some problems, the uniform allocation strategy seems to be the best one, and this has to appear somewhere in the upper bounds. On the other hand, consider a case when there are several suboptimal arms, a few of them with small gaps Δ_j, all other ones with large gaps Δ_j. One can take advantage of this situation with an adaptive algorithm such as UCB(p) and quickly focus on a small subset of good candidates among the arms. For similar problems, the bound on the simple regret for this forecaster should be better than the one corresponding to the uniform allocation strategy, see Section 3.4. Note also that in all cases, as discussed in details in Section 4, the simple regret of the uniform allocation strategy will be smaller than the one of UCB(p) after some time. These subtleties lead to a less easy analysis than the classical one for the cumulative regret.

3.1 Overview of the bounds

Table 1 summarizes the distribution-dependent and distribution-free bounds we could prove so far. It shows that two interesting couple of strategies are, on one hand, the uniform allocation together with the choice of the empirical best arm, and on the other hand, UCB(p) together with the choice of the most played arm. The first pair was perhaps expected, the second one might be considered more suprising.

We only prove here upper bounds on the simple regrets of these two pairs and omit the proofs of all other upper bounds. The distribution-dependent lower bound is stated in Corollary 2 and the distribution-free lower bound follows from a straightforward adaptation of the proof of the lower bound on the cumulative regret in [ACBFS02].

Table 1 indicates that while for distribution-dependent bounds, the optimal rates of decrease in the number n of rounds for simple regrets is exponential, for distribution-free bounds, the rate worsens to $1/\sqrt{n}$. A similar situation arises for the cumulative regret, see [LR85] (optimal $\ln n$ rate for distribution-dependent bounds) versus [ACBFS02] (optimal $\sqrt{n}$ rate for distribution-free bounds).

	Distribution-dependent		
	EDP	EBA	MPA
Uniform		$\bigcirc\, e^{-\bigcirc n}$	
UCB(p)	$\bigcirc (p \ln n)/n$	$\bigcirc\, n^{-\bigcirc}$	$\bigcirc\, n^{2(1-p)}$
Lower bound		$\bigcirc\, e^{-\bigcirc n}$	

	Distribution-free		
	EDP	EBA	MPA
Uniform		$\square \sqrt{\frac{K \ln K}{n}}$	
UCB(p)	$\square \sqrt{\frac{pK \ln n}{n}}$	$\frac{\square}{\sqrt{p \ln n}}$	$\square \sqrt{\frac{pK \ln n}{n}}$
Lower bound		$\square \sqrt{\frac{K}{n}}$	

Table 1: Distribution-dependent (top) and distribution-free (bottom) bounds on the expected simple regret of the considered pairs of allocation (lines) and recommendation (columns) strategies. Lower bounds are also indicated. The $\square$ symbols denote the universal constants, whereas the $\bigcirc$ are distribution-dependent constants.

3.2 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: it achieves the optimal rates of decrease both in terms of its distribution-dependent and distribution-free bounds.

Below, we mean by the recommendation given by the empirical best arm at round $K\lfloor n/K \rfloor$ the recommendation $\psi_{K\lfloor n/K \rfloor}$ of EBA (see Figure 3), where $\lfloor x \rfloor$ denotes the lower integer part of a real number x. The reason why we prefer $\psi_{K\lfloor n/K \rfloor}$ to ψ_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 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 n. Both match the lower bound exhibited in Corollary 2.

Proposition 1 *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round $K\lfloor n/K \rfloor$) ensures that the simple regrets are bounded by*

$$\mathbb{E} r_n \leqslant \sum_{j:\Delta_j>0} \Delta_j \, e^{-\Delta_j^2 \lfloor n/K \rfloor/2}$$

for all $n \geqslant K$; and by

$$\mathbb{E} r_n \leqslant \left(\max_{j=1,\dots,K} \Delta_j \right) \exp\left(-\frac{1}{8} \left\lfloor \frac{n}{K} \right\rfloor \Delta^2 \right)$$

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,

$$\mathbb{E} r_n = \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}\psi_{j,n} \leqslant \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{P}\big\{ \widehat{\mu}_{j,n} \geqslant \widehat{\mu}_{j^*,n} \big\}$$

where the upper bound follows from the fact that to be the empirical best arm, an arm j must have performed, in particular, better than a best arm j^*. We now apply Hoeffding's inequality (for i.i.d. random variables, see [Hoe63]). $\widehat{\mu}_{j,n} - \widehat{\mu}_{j^*,n}$ is an average of $\lfloor n/K \rfloor$ i.i.d. random variables bounded between -1 and 1 and with common expectation $-\Delta_j$. Thus, the probability of interest is bounded by

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

which yields the first result.

The second inequality is proved by resorting to a sharper concentration argument, namely, the method of bounded differences, see [McD89], see also [DL01, Chapter 2]. The proof, less central to this paper, will be eventually omitted and can be found, for now, in appendix. ■

The distribution-free bound is obtained not as a corollary of the distribution-dependent bound, but as a consequence of its proof. A direct optimization over the Δ_j in the first bound of Proposition 1 indeed yields a suboptimal distribution-free bound. One therefore has to proceed with slightly more care.

Corollary 4 *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round*

$K\lfloor n/K\rfloor$) ensures that the simple regrets are bounded in a distribution-free sense, for $n \geqslant K$, as

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

Proof: It is not enough to optimize the bound of Proposition 1 over the Δ_j, for it would yield an additional multiplicative factor of K. Instead, we extract from its proof that

$$\mathbb{E}\psi_{j,n} \leqslant \exp\left(-\frac{1}{2}\left\lfloor\frac{n}{K}\right\rfloor\Delta_j^2\right)\,;$$

we now distinguish whether a given Δ_j is more or less than a threshold ε, use that $\sum\psi_{j,n}=1$ and $\Delta_j\leqslant 1$ for all j, and can thus write

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

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. ■

3.3 Analysis of UCB(p) combined with MPA

We need a technical lemma and then exploit it to obtain two different distribution-dependent bounds of different utilities.

Lemma 5 *The allocation strategy given by UCB(p) (where $p>1$) associated to the recommendation given by the most played arm ensures that the simple regrets are bounded in a distribution-dependent sense as follows. For all $a_1,\ldots,a_K$ such that $a_j\geqslant 0$ for all j, with $a_1+\ldots+a_K=1$, and such that for all suboptimal arms j and all optimal arms j^*, one has $a_j\leqslant a_{j^*}$,*

$$\mathbb{E}r_n \leqslant \frac{1}{p-1}\sum_{j\neq j^*}(a_j n)^{2(1-p)}$$

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

$$a_j n \geqslant 1+\frac{4p\ln n}{\Delta_j^2} \quad \textit{and} \quad a_j n\geqslant K+2\,.$$

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

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

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

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

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

Thus, using that $\Delta_j\leqslant 1$ for all j, the simple regret can be bounded as

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n^*} \leqslant \sum_{j:\Delta_j>0}\mathbb{P}\big\{T_j(n)\geqslant a_j n\big\}\,.$$

A side-result extracted from the proof of [ACBF02, Theorem 1] states that for all suboptimal arms j and all rounds $t\geqslant K+1$,

$$\mathbb{P}\big\{I_t=j \text{ and } T_j(t-1)\geqslant\ell\big\}\leqslant 2\,t^{1-2p} \qquad \text{whenever} \qquad \ell\geqslant\frac{4p\ln n}{\Delta_j^2}. \quad (3)$$

This yields that for a suboptimal arm j and since by the assumptions on n and the a_j, the choice $\ell=a_j n-1$ satisfies $\ell\geqslant K+1$ and $\ell\geqslant(4p\ln n)/\Delta_j^2$,

$$\begin{aligned}
&\mathbb{P}\big\{T_j(n)\geqslant a_j n\big\}\\
&\quad\leqslant \sum_{t=a_j n}^{n}\mathbb{P}\Big\{T_j(t-1)=a_j n-1 \text{ and } I_t=j\Big\}\\
&\quad\leqslant \sum_{t=a_j n}^{n}2\,t^{1-2p}\leqslant\frac{1}{p-1}(a_j n)^{2(1-p)} \qquad (4)
\end{aligned}$$

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

A first distribution-dependent bound is stated below; the bound does not involve any quantity depending on the Δ_j, but it only holds for rounds n large enough, a statement that does involve the Δ_j. Its interest is first, that its bound is simple to read, and second, that the techniques used to prove it imply easily a distribution-free bound, stated in Theorem 7 and which is comparable to Corollary 4. A discussion of the earlier results of [KS06] follows, as well as a second distribution-dependent bound that will be compared to Proposition 1 in Section 3.4.

Theorem 6 *The allocation strategy given by UCB(p) (where $p>1$) associated to the recommendation given by the most played arm ensures that the simple regrets are bounded in a distribution-dependent sense by*

$$\mathbb{E}r_n\leqslant\frac{K^{2p-1}}{p-1}\,n^{2(1-p)}$$

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

$$n\geqslant K+\frac{4Kp\ln n}{\Delta^2} \quad \textit{and} \quad n\geqslant K(K+2)\,.$$

This result matches the lower bound exhibited in Corollary 3; in the upper bound presented above, the polynomial rate of decrease is distribution-free. In addition, it illustrates Theorem 1: the larger p, the larger the (theoretically guaranteed bound on the) cumulative regret of UCB(p) but the smaller the simple regret of UCB(p) associated to the recommendation given by the most played arm.

Proof: We apply Lemma 5 with the uniform choice $a_j = 1/K$ and recall that Δ is the minimum of the $\Delta_j > 0$. ■

Theorem 7 *The allocation strategy given by UCB(p) (where $p > 1$) associated to the recommendation given by the most played arm ensures that the simple regrets are bounded in a distribution-free sense by*

$$\begin{aligned}\mathbb{E}r_n &\leqslant \sqrt{\frac{4Kp\ln n}{n-K}} + \frac{K^{2p-1}}{p-1}\, n^{2(1-p)} \\ &= O\left(\sqrt{\frac{Kp\ln n}{n}}\right)\end{aligned}$$

for all $n \geqslant K(K+2)$.

Proof: We start the proof by applying (2), which holds in general, as well as the fact that $J_n^* = j$ only if $T_j(n) \geqslant n/K$, that is,

$$\psi_{j,n} = \mathbb{I}_{\{J_n^*=j\}} \leqslant \mathbb{I}_{\{T_j(n)\geqslant n/K\}}\ ,$$

to get

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n^*} \leqslant \varepsilon + \sum_{j:\Delta_j>\varepsilon} \Delta_j\, \mathbb{P}\left\{T_j(n) \geqslant \frac{n}{K}\right\}\ .$$

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

$$\mathbb{E}r_n \leqslant \varepsilon + \sum_{j:\Delta_j>\varepsilon} \frac{\Delta_j}{p-1}\, K^{2(p-1)}\, n^{2(1-p)}$$

where ε is chosen such that for all $\Delta_j > \varepsilon$, the needed condition $\ell = n/K - 1 \geqslant (4p\ln n)/\Delta_j^2$ 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,

$$\varepsilon = \sqrt{\frac{4pK\ln n}{n-K}}\ ,$$

and upper bounding all remaining Δ_j by 1. ■

Remark 1 We can rephrase the results of [KS06] as using UCB1 as an allocation strategy and forming a recommendation according to the empirical best arm. In particular, [KS06, Theorem 5] provides a distribution-dependent bound on the probability of not picking the best arm with this procedure. It can be used to derive a bound on the simple regret. By reproducing their calculations (to have an explicit expression of the leading constant in terms of the Δ_j), we got

$$\mathbb{E}r_n \leqslant \sum_{j:\Delta_j>0} \frac{4}{\Delta_j}\left(\frac{1}{n}\right)^{\rho\Delta_j^2/2}$$

for all $n \geqslant 1$. This bound, in particular because of the leading constants $1/\Delta_j$ and of the distribution-dependant exponent, is not as nice as the bound presented in Theorem 6. The best distribution-free bound we could get from this bound was of the order of $1/\sqrt{\ln n}$, a rate by far slower than the optimal $1/\sqrt{n}$ rate stated in Theorem 7.

3.4 Discussion: Comparison of the bounds

We now explain why, in some cases, the bound provided by our theoretical analysis in Lemma 5 is better than the bound stated in Proposition 1. This will be further illustrated in the simulation section.

The central point in the argument is that the bound of Lemma 5 is of the form $\bigcirc\, n^{2(1-p)}$, 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_j in the rate of convergence.

Some care is needed in the comparison, since the bound for UCB(p) 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(p) 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 $\lfloor n/K \rfloor$ each arm, whereas the UCB strategy focuses rapidly its exploration on the best arms.

A general detailed argument to provide such examples is provided in the appendix. It will be omitted in the final version, where only the heuristic arguments above will be kept.

4 A brief simulation study

We propose three 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 regrets). The first one corresponds in some sense to the worst case alluded at at the beginning of Section 3. It shows that for small values of n (e.g., $n \leqslant 80$ in Figure 4), the uniform allocation strategy is very competitive. Of course the range of these values of n can be made arbitrarily large by decreasing the gaps. The second one corresponds to the discussion in Section 3.4, while the third one represents a rather typical behavior of the strategies when K is large.

The attentive reader may be surprised that we never see the uniform allocation strategy converging more rapidly than UCB–based strategies, whereas the combination of the lower bound of Corollary 3 and the upper bound of Proposition 1 shows that for all distributions over the arms, the uniform allocation strategy will be better than UCB(p) after some point n. Actually, this n is very large, so large that at it, the simple regrets are already below computers precision. 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.

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

This section is of theoretical interest. We consider the $\mathcal{X}$-armed bandit problem of, e.g., [Kle04, BMSS09] and (re-)define the notions of cumulative and simple regrets. 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 of its natural interpretation, the simple regret thus appears as a tool for proving results on the cumulative regret.

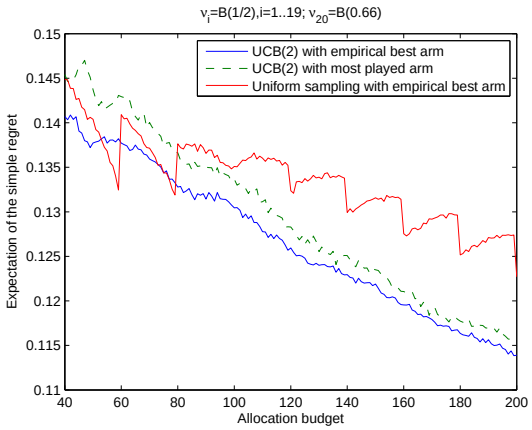

Figure 4: $K = 20$ arms with Bernoulli distributions of parameters 0.50 for the first 19th of them and 0.66 for the last one.

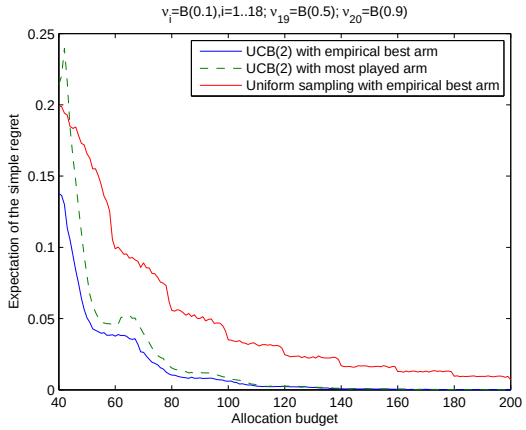

Figure 5: Parameters of the numerical application in Section B.

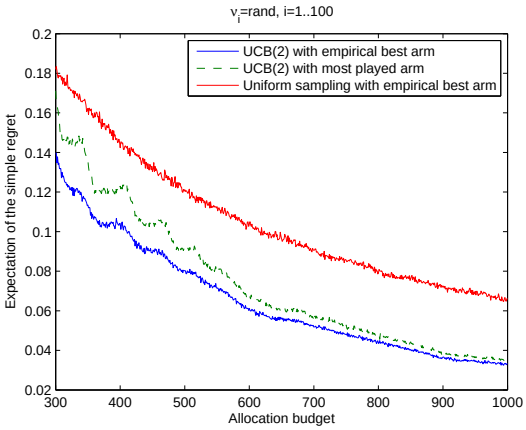

Figure 6: $K = 100$ arms with Bernoulli distributions, whose parameters are chosen independently at random in $[0, 1]$.

Parameters: an environment E on $\mathcal{X}$

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 $E(I_t)$;

Aim:
Find a pulling 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., $o(n)$).

Figure 7: $\mathcal{X}$–armed bandit problems.

5.1 Description of $\mathcal{X}$–armed bandit problems

For a (bounded) interval B 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 7 and 8. There, an environment E on $\mathcal{X}$ is fixed by Nature and we want various notions of regret to be small, given this environment.

We consider now families of environments and say that a family $\mathcal{F}$ of environments is explorable–exploitable (respectively, explorable) if there exists a strategy 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 2.

But surprisingly enough, 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 9 are simpler to exhibit and prove.

Lemma 8 *A family of environments $\mathcal{F}$ is explorable if and only if it is explorable–exploitable.*

The proof will be omitted from this extended abstract and can be found in the appendix. 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 exploitation are fed into the base allocation strategy.

5.2 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 $\mathcal{C}\left(\mathcal{P}([0, 1])^{\mathcal{X}}\right)$ the subset of $\mathcal{P}([0, 1])^{\mathcal{X}}$

Parameters: environment E on $\mathcal{X}$

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.

Aim:
Find an allocation strategy (φ_t) and a recommendation strategy (ψ_n) 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., $o(1)$).

Figure 8: The pure exploration problem for $\mathcal{X}$–armed bandit problems.

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 9 *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 10 *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 the appendix. 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.

6 Conclusions and future work

We introduced a notion of simple regret, that models the situations where a forecaster is given an exploration phase before outputting a recommendation. We showed that the exploration–exploitation trade-off needed to minimize the simple regret is quantitatively different from the one to be used when minimizing the cumulative regret. We provided distribution-dependent and distribution-free bounds on the simple regret. As long as distribution-dependent upper bounds are concerned, asymptotic behaviors are in favor of the uniform (or linear) exploration of each arm. However, as illustrated by the simulations, this asymptotic phase seems to occur when computer precision limits are reached and UCB–based strategies perform better than the uniform allocation for moderate values of n. We provided situations where the superiority of UCB–based strategies over the uniform allocation is reflected in the bounds, and we believe this line of analysis may be extended. Possible directions include improving the analysis of the performance of UCB–based strategies strategies both for the simple and cumulative regrets.

References

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47:235–256, 2002.

[ACBFS02] 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, 2002.

[Bil68] P. Billingsley. *Convergence of Probability Measures*. Wiley and Sons, 1968.

[BMSS09] Sébastien Bubeck, Remi Munos, Gilles Stoltz, and Csaba Szepesvari. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems 21*, 2009.

[CM07] P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[EDMM02] 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.

[GWMT06] S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in Monte-Carlo go. Technical Report RR-6062, INRIA, 2006.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Kle04] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems*, 2004.

[KS06] L. Kocsis and Cs. Szepesvari. Bandit based Monte-carlo planning. In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.

[LR85] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics (Proceedings of the 12th British Combinatorial Conference)*, pages 148–188, 1989.

[MLG04] O. Madani, D. Lizotte, and R. Greiner. The budgeted multi-armed bandit problem. pages 643–645, 2004. Open problems session.

[MT04] 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.

[Rob52] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

[Sch06] K. Schlag. Eleven tests needed for a recommendation. Technical Report ECO2006/2, European University Institute, 2006.

A Proofs and discussion to be omitted from the final version

A.1 Proof of the second statement of Proposition 1

Proof: We start by writing

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

where the inequality follows from the fact that regret is suffered only when an arm with suboptimal expectation has the best empirical performances. Now, the quantity of interest can be rewritten as

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

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 [McD89], see also [DL01, 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

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

for all $\varepsilon > 0$. We choose

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

(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

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

this is done by a classical argument, using bounds on the moment generating function of the random variables of interest. Consider $Z_j = \lfloor n/K \rfloor \left(\widehat{\mu}_{j,n} - \widehat{\mu}_{j^*,n} + \Delta_j\right)$ for all $j = 1, \dots, K$. Independence and Hoeffding's lemma (see, e.g., [DL01, Chapter 2]) imply that for all $\lambda > 0$,

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

(where we used again that Z_j 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, [DL01, Chapter 2]) then yields

$$\mathbb{E}\left[\max_{j=1,\dots,K} Z_j\right] \leqslant \sqrt{2\lfloor n/K \rfloor \ln K} \, ,$$

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

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

(for n sufficiently large, a statement made precise below), one has

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

The result follows for n such that

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

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

B Detailed discussion of the heuristic arguments presented in Section 3.4

We first state the following corollary to Lemma 5.

Theorem 11 *The allocation strategy given by UCB(p) (where $p > 1$) associated to the recommendation given by the most played arm ensures that the simple regrets are bounded in a distribution-dependent sense by*

$$\mathbb{E}r_n \leqslant \frac{n^{2(1-p)}}{p-1} \sum_{j \neq j^*} \left(\frac{\Delta_j^2}{\beta}\right)^{2(p-1)}$$

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

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

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

$$\beta = \frac{1}{\dfrac{K^*}{\Delta^2} + \displaystyle\sum_{j \neq j^*} \frac{1}{\Delta_j^2}} \, .$$

Proof: We apply Lemma 5 with the choice $a_j = \beta/\Delta_j^2$ for all suboptimal arms j and $a_{j^*} = \beta/\Delta^2$ for all optimal arms j^*, 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 11 implies that

$$\mathbb{E}r_n \leqslant \frac{K-1}{p-1}\left(\frac{6+K}{n}\right)^{2(p-1)} \tag{5}$$

for all n sufficiently large, e.g.,

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

Now, the upper bound on $\mathbb{E}r_n$ for the uniform allocation given in Proposition 1 is larger than

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

Thus for n such that

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

and $(4p+1)\ln n \geqslant (K+2)4\Delta^2$, the bound for the uniform allocation is at least

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

which may be much worse than the upper bound (5) for the UCB(p) strategy, whenever K is large, as can be seen by comparing the exponents $-2(p-1)$ versus $-(4p+1)(6+K)/8K$.

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 $p = 4$. Then $n = 6020$ satisfies (6) and the upper bound (5) for the UCB(p) strategy is 4.11×10^{-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.

C Proof of Lemma 8

Proof: In view of the comments before the statement of Lemma 8, 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_n) strategies such that for all environments $E \in \mathcal{F}$, the simple regrets $\mathbb{E}r_n = o(1)$, and provide a new strategy (φ'_t) such that its cumulative regret $\mathbb{E}R'_n = o(n)$ for all environments E.

It is defined informally as follows. At round $t = 1$, it uses $\varphi'_1 = \varphi_1$ and gets a reward $Y'_1 = Y_1$. Based on this reward, the recommendation ψ_1 is formed and at round $t = 2$, the new strategy plays $\varphi'_2 = \psi_1$. It gets a reward but does not take it into account. It bases its choice $\varphi'_3 = \varphi_2$ only on Y'_1, and gets a reward $Y'_2 = Y_3$. Based on Y'_1 and Y'_2, the recommendation ψ_2 is formed and played at rounds $t = 4$ and $t = 5$, i.e., $\varphi'_4 = \varphi'_5 = \psi_2$. And so on: the sequence of distributions chosen by the new strategy is given by

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

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

$$n = \frac{t(n)\big(t(n)+1\big)}{2} + k(n) \qquad \text{where } k(n) \in \big\{0,\dots,t(n)\big\}. \tag{7}$$

Then,

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

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, i.e.,

$$r_s = \sup_{x\in\mathcal{X}}\mu(x) - \int_{\mathcal{X}}\mu(x)\,\mathrm{d}\psi_s(x).$$

Taking expectations with respect to any fixed environment, we get

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

the first term in the right-hand side is of the order of $1/\sqrt{n}$ and the second one is a Cesaro average and thus converges to 0. This concludes that the exhibited strategy has a small cumulative regret for all environments of the family, which is thus explorable–exploitable. ■

D Proof of Theorem 9 and its corollary

The key ingredient is the following characterization of separability (which relies on an application of Zorn's lemma); see, e.g., [Bil68, Appendix I, page 216].

Lemma 12 *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*

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

A nice application (which we do however not fully need in the proof of Theorem 9, we only use the straightforward direct part) is the following characterization of separability 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.

Lemma 13 *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 $x_1, x_2, \dots$ a dense sequence. If it is finite with length N, we let

$$\lambda = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_j}$$

and otherwise,

$$\lambda = \sum_{j \geqslant 1} \frac{1}{2^j} \delta_{x_j} \; .$$

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

For the direct implication, we use Lemma 12 (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. ∎

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

The proof 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 13.

The proposed strategy is then constructed in a way similar to the one exhibited in Section C, 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). Note that exploration and exploitation appear in two distinct phases, as was the case already, for instance, in Section 3.2.

We now denote by

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

the simple regret can then be decomposed as

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

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

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

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

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

since this holds for all $\varepsilon > 0$, we have the desired convergence

$$\mu^* - \mathbb{E}\Big[\mu^*_{g(n)}\Big] \longrightarrow 0 \; .$$

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

$$\begin{aligned} &\mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big] \qquad (8) \\ &\leqslant \; 2\varepsilon \\ &\quad + \mathbb{P}\Big\{\exists \, s \leqslant g(n), \;\; \Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon\Big\} \; . \end{aligned}$$

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

$$\mathbb{P}\Big\{\Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon \;\Big|\; \mathcal{A}_n\Big\} \qquad (9)$$

for $1 \leqslant s \leqslant g(n)$, where $\mathcal{A}_n$ is the σ–algebra generated by 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 (9) are bounded, via an application of Hoeffding's inequality (for i.i.d. random variables, see [Hoe63]), 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 $s+1,\ldots,t(n)-1$, and maybe in $t(n)$, where n is decomposed again as in (7). 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 (7) implies

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

Substituting this in the Hoeffding's bound, integrating, taking a union bound leads from (8) 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

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

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

D.2 Separability of $\mathcal{X}$ is a necessary condition

This basically follows from the impossibility of a uniform exploration, as asserted by Lemma 13.

Proof: Let $\mathcal{X}$ be a non-separable metric space (with distance denoted by d). Let A be an uncountable set and $\rho > 0$ defined as in Lemma 12; 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

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

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 $x \in \mathcal{X} \setminus B(a,\rho/2)$, and 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 $n \geqslant 1$,

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

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 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 $\varphi_1 = \varphi^0_1$ over $\mathcal{X}$, picks I_1 at random according to φ^0_1, gets a deterministic payoff $Y_1 = 0$, and finally recommends $\psi^0_1(I_1) = \psi_1(I_1, Y_1)$ (which depends on I_1 only, since the obtained payoffs are all null). In the second round, it chooses an allocation $\psi^0_2(I_1)$ (that depends only on I_1, for the same reasons as before), picks I_2 at random according to $\psi^0_2(I_1)$, gets a null reward, and recommends $\psi^0_2(I_1, I_2)$; 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 measurable applications

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

we introduce the distribution $\mathbb{A}\cdot\nu \in \mathcal{P}(\mathcal{X})$ 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, $\mathbb{A}\cdot\varphi^0_t$ and $\mathbb{A}\cdot\psi^0_t$ give a positive probability mass to $B(a,\rho/2)$; and 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 $n \geqslant 1$,

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

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 had to be proved. ∎

D.3 The countable case of Corollary 10

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 9. ∎

D.4 An additional remark

Remark 2 In this extended abstract, we only consider non-uniform bounds. Uniform bounds, i.e., bounds for

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

can be exhibited 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.