Pure Exploration in Multi-Armed Bandits Problems

Sébastien Bubeck[1], Rémi Munos[1], and Gilles Stoltz[2,3]

[1] INRIA Lille, SequeL Project, France
[2] Ecole normale supérieure, CNRS, Paris, France
[3] HEC Paris, CNRS, Jouy-en-Josas, France

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.

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 resources (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. Unlike the latter, this criterion is not suited for a distribution-free analysis.

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 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.

Fig. 1. The pure exploration problem for multi-armed bandits.

2 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 r_t 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,

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

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, 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., [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 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. It in particular 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.

3 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, that is, $\psi_n = (\delta_{I_1} + \ldots + \delta_{I_n})/n$, 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" trade-off 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 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.

Theorem 1 (Main result). *For all allocation strategies (φ_t) and all functions $\varepsilon : \{1, 2, \dots\} \to \mathbb{R}$ 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 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 $D \geqslant 0$ and an ordering $\nu_1, \dots, \nu_K$ of the considered distributions with*
>
> $$\mathbb{E}r_n \geqslant \frac{\Delta}{2}\, e^{-D\varepsilon(n)} .$$

Corollary 1. *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 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 regrets are 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 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 2. *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 $\gamma \geqslant 0$ (independent of p) such that, up to the choice of a good ordering of the considered distributions,*

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

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 ordering) such that $1 > \mu_1 > \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_K \geqslant 0$ and $\mu_2 > \mu_K$ (thus, $\mu_2 > 0$).

Another layer of notation is needed. It 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 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 reordering. $\mathbb{P}_\sigma$ 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 reordering,

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

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

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

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

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

and therefore,

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

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

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

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 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

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

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 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,

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

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 $\sigma(2),\dots,\sigma(K-1)$, we denote by $\mathbb{P}^{\sigma(2),\dots,\sigma(K-1)}$ 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 $\{1,\dots,K\}$ not present in the $(K-2)$–tuple, and the other one being such that $\sigma(1)$ equals the other such element. Formally,

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

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

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

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

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 the times in the exploration phase can have interesting simple regrets. 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. The second one is UCB(p) (a variant of UCB1 where the quantile factor may be a parameter); it is designed for the classical exploration–exploitation dilemma (i.e., its 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 cumulative and simple regrets, we restrict our attention to the two discussed above.

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 Figures 2 and 3.

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 surprising. We only state

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

(2) use $\varphi_t = \delta_{j^*_{t-1}}$, where $j^*_{t-1} \in \underset{j=1,\ldots,K}{\operatorname{argmax}} \widehat{\mu}_{j,t-1} + \sqrt{\dfrac{p \ln(t-1)}{T_j(t-1)}}$
(ties broken by choosing, for instance, the arm with smallest index).

Fig. 2. Two allocation strategies.

here upper bounds on the simple regrets of these two pairs and omit the other ones. The distribution-dependent lower bound is stated in Corollary 1 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 asymptotic optimal rate 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).

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 states its theoretical properties: the rate of decrease of its simple regret is exponential in a distribution-dependent sense and equals the optimal (up to a logarithmic term) $1/\sqrt{n}$ rate in the distribution-free case. In Proposition 1, 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. Their simple proof is omitted; it relies on concentration inequalities, namely, Hoeffding's inequality and McDiarmid's inequality. The distribution-free bound of Corollary 3 is obtained not as a corollary of Proposition 1, but as a consequence of its proof. Its simple proof is also omitted.

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

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),

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

(ties broken in some way).

Empirical distribution of plays (EDP)

Draws a recommendation using the probability distribution $\psi_n = \dfrac{1}{n} \displaystyle\sum_{t=1}^{n} \delta_{I_t}$.

Fig. 3. Three recommendation strategies.

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

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.

Proposition 1. *The uniform allocation strategy associated to the recommendation given by the empirical best arm ensures that the simple regrets are bounded as follows:*

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

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

Corollary 3. *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*

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

4.2 Analysis of UCB(p) combined with MPA

A first (distribution-dependent) bound is stated in Theorem 2; 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 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. *For $p > 1$, the allocation strategy given by UCB(p) 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 $n \geqslant K + \dfrac{4Kp\ln n}{\Delta^2}$ 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 p, the larger the (theoretical 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.

Theorem 3. *For $p > 1$, the allocation strategy given by UCB(p) associated to the recommendation given by the most played arm ensures that the simple regrets are bounded for all $n \geqslant K(K+2)$ in a distribution-free sense by*

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

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 and can be used to derive the following bound on the simple regret:

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

for all $n \geqslant 1$. The leading constants $1/\Delta_j$ and the distribution-dependant 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 $1/\sqrt{\ln n}$, to be compared to the asymptotic optimal $1/\sqrt{n}$ rate stated in Theorem 3.

Proofs of Theorems 2 and 3

Lemma 1. *For $p > 1$, the allocation strategy given by UCB(p) 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 $a_1 + \ldots + a_K = 1$ and $a_j \geqslant 0$ for all j, with the additional property that for all suboptimal arms j and all optimal arms j^*, one has $a_j \leqslant a_{j^*}$, the following bound holds:*

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

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 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,

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

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

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

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 $\ell \geqslant (4p\ln n)/\Delta_j^2$,

$$\begin{aligned}\mathbb{P}\big\{T_j(n) \geqslant a_j n\big\} &\leqslant \sum_{t=a_j n}^{n} \mathbb{P}\Big\{T_j(t-1) = a_j n - 1 \ \text{ and } \ I_t = j\Big\} \\ &\leqslant \sum_{t=a_j n}^{n} 2\, t^{1-2p} \leqslant \frac{1}{p-1}(a_j n)^{2(1-p)} \end{aligned} \tag{3}$$

where we used a union bound for the second inequality and (2) for the third inequality. A summation over all suboptimal arms j concludes the proof.

Proof (of Theorem 2). We apply Lemma 1 with the uniform choice $a_j = 1/K$ and recall that Δ is the minimum of the $\Delta_j > 0$.

Proof (of Theorem 3). We start the proof by using that $\sum \psi_{j,n} = 1$ and $\Delta_j \leqslant 1$ for all j, and can thus write

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n^*} = \sum_{j=1}^{K} \Delta_j\, \mathbb{E}\psi_{j,n} \leqslant \varepsilon + \sum_{j:\Delta_j > \varepsilon} \Delta_j\, \mathbb{E}\psi_{j,n}\ .$$

Since $J_n^* = j$ only if $T_j(n) \geqslant n/K$, that is, $\psi_{j,n} = \mathbb{I}_{\{J_n^*=j\}} \leqslant \mathbb{I}_{\{T_j(n)\geqslant n/K\}}$, we get

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

Applying (3) with $a_j = 1/K$ leads to $\displaystyle \mathbb{E}r_n \leqslant \varepsilon + \sum_{j:\Delta_j>\varepsilon} \frac{\Delta_j}{p-1}\, K^{2(p-1)}\, n^{2(1-p)}$

where ε is chosen such that for all $\Delta_j > \varepsilon$, the 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, $\varepsilon = \sqrt{(4pK\ln n)/(n-K)}$ and upper bounding all remaining Δ_j by 1.

5 Conclusions: Comparison of the bounds, simulation study

We now explain why, in some cases, the bound provided by our theoretical analysis in Lemma 1 is better than the bound stated in Proposition 1. The central point in the argument is that the bound of Lemma 1 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

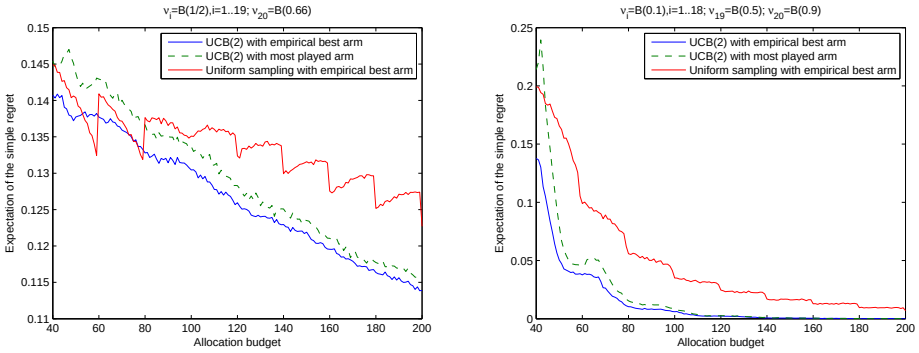

Fig. 4. $K = 20$ arms with Bernoulli distributions of parameters indicated on top of each graph.

focuses rapidly its exploration on the best arms. A general argument is proposed in the extended version [BMS09, Appendix B]. We only consider here one numerical example extracted from there, see the right part of Figure 4. For moderate values of n (at least when n is about $6\,000$), the bounds associated to the sampling with UCB(p) are better than the ones associated to the uniform sampling.

To make the story described in this paper short, we can distinguish three regimes:

- 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(p) is preferable, as discussed just above;
- for small values of n, the best bounds to use seem to be the distribution-free bounds, which are of the same order of magnitude for the two strategies.

Of course, these statements involve distribution-dependent quantifications (to determine which n are small, moderate, or large).

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 regrets. (More experiments can be found in [BMS09].) The first one corresponds in some sense to the worst case alluded at the beginning of Section 4. It shows that for small values of n (e.g., $n \leqslant 80$ in the left plot of 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 numerical example described earlier in this section.

We mostly illustrate here the small and moderate n regimes. Because of the chosen ranges, we do not see yet the uniform allocation strategy getting better than UCB–based strategies. This is because for large n, the simple regrets are usually very small, even below computer 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 only.

6 Pure exploration for bandit problems in topological spaces

These results are of theoretical interest. We summarize them very briefly; statements and proofs can be found in the extended version [BMS09]. Therein, we consider the $\mathcal{X}$–armed bandit problem with bounded payoffs of, e.g., [Kle04,BMSS09] and (re-)define the notions of cumulative and simple regrets. The topological set $\mathcal{X}$ is a large possibly non-parametric space but the associated mean-payoff function is continuous. We show that, without any assumption on $\mathcal{X}$, there exists a strategy with cumulative regret $\mathbb{E}R_n = o(n)$ if and only if there exist an allocation and a recommendation strategy with simple regret $\mathbb{E}r_n = o(1)$. We then use this equivalence to characterize the metric spaces $\mathcal{X}$ in which the cumulative regret $\mathbb{E}R_n$ can always be made $o(n)$: they are given by the separable spaces. Thus, here, in addition of its natural interpretation, the simple regret appears as a tool for proving results on the cumulative regret.

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.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration for multi-armed bandit problems. Technical report, HAL report hal-00257454, 2009. Available at http://hal.archives-ouvertes.fr/hal-00257454/en.

[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.

[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.

[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.

[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.