Pure Exploration for Multi-Armed Bandit Problems

Sébastien Bubeck, Rémi Munos, Gilles Stoltz

Pure Exploration for Multi-Armed Bandit Problems

Sébastien Bubeck
SequeL Project, INRIA Futurs Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France
sebastien.bubeck@inria.fr

Rémi Munos
SequeL Project, INRIA Futurs Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France
remi.munos@inria.fr

Gilles Stoltz*
Ecole Normale Supérieure, CNRS
75005 Paris, France
HEC Paris School of Management, CNRS,
78351 Jouy-en-Josas, 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 new 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. Our goal is to demonstrate that quite counter-intuitively, exploration–exploitation trade-offs are still valuable in this setting. We do so by first providing an experimental study and then aim at explaining theoretically the observed phenomenon. A first negative result is that too small a cumulative regret prevents the simple regret from decreasing exponentially towards zero, its optimal distribution-dependent rate. We solve the paradox by considering distribution-free bounds and pointing out two regimes for distribution-dependent bounds.

1 Introduction and motivation

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 Robbins (1952), a sequential decision problem where, at each stage, the forecaster has to pull one 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 difference between the expected reward of the best arm and the average obtained rewards. Typical good strategies, like the UCB strategies of Auer et al. (2002a), trade off between exploration and exploitation.

*Partially supported by the French "Agence Nationale pour la Recherche" under grant JCJC06-137444 "From applications to theory in learning and adaptive statistics" and by the PASCAL Network of Excellence under EC grant no. 506778.

When exploration involves costs not measured in terms of rewards but rather in terms of numerical resources (e.g., memory or CPU), the performances of a strategy have to be assessed in a different way. The forecaster may then allocate sequentially its resources to explore the arms to his convenience and output the index of an arm when these resources have been all used. The recommended arm is assessed on a new one-shot instance of the same bandit problem, leading to the notion of simple regret: the (expectation of the) difference between the reward of the best arm and the one of the recommended arm. We term this variant the pure exploration problem, since, at first sight, exploration and exploitation appear in two distinct phases (a statement we shall however qualify later on).

A concrete example for such a pure exploration problem is given by tree search, for which strategies minimizing the cumulative regret have been used recently in a hierarchical way to guarantee an exploration making a good use of available CPU time. Namely, the UCT strategy of Kocsis and Szepesvari (2006) and the BAST strategy of Coquelin and Munos (2007) have shown interesting performances for solving minimax tree search problems with huge trees; they have been applied successfully to the game of go, see, for instance, the MoGo program of Gelly et al. (2006) that plays at a world-class level. The tree exploration policy resulted in an asymmetric tree expansion in which the most promising edges were explored first. Strategies designed to minimize the simple regret (instead of the cumulative regret) are expected to be the stone for an improvement of these results.

This pure exploration problem was referred to as "budgeted multi-armed bandit problem" in the open problem by Madani et al. (2004). Schlag (2006) solves the pure exploration problem in a minmax sense for the case of two arms only and rewards given by probability distributions over $[0, 1]$. It has also been studied in related settings. Even-Dar et al. (2002) and Mannor and Tsitsiklis (2004) consider forecasters performing exploration during a random number of rounds T and aiming 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. However, the algorithms proposed in the references above do not come with anytime performances yet, which we think would be necessary (but is not straightforward if done without a doubling trick). If, for instance, the forecaster is unsure about the available computational power of, e.g., a shared system he is using, and is given a fixed delay to perform exploration and recommend an arm, he cannot guess the number of available rounds. For non-crucial medical applications where the regret would be a suitable measure of efficiency (e.g., test of slimming pills), a test phase may last a given time but one cannot determine in advance how many patients will be included in the study. As a consequence, we do not assume the knowledge of available rounds in the sequel.

2 Problem setup, notation

We consider a sequential decision problem for multi-armed bandits, where a forecaster plays against a stochastic environment. $K \geq 2$ arms, denoted by $j = 1, \ldots, K$, are available and the j-th of them is parameterized by a probability distribution ν_j (with finite first moment and 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 $t \geq 1$, we denote by $N_{j,t}$ the number of times j was pulled from rounds 1 to t, and by $X_{j,1}, X_{j,2}, \ldots, X_{j,N_{j,t}}$ the sequence of associated rewards.

The forecaster has to deal simultaneously with two tasks, a primary one and an auxiliary one. The auxiliary task consists in exploration, the forecaster should indicate at each round t the arm I_t to be pulled. He may resort to a randomized strategy, denoted by $\varphi_t \in \Delta\{1, \ldots, K\}$ (where $\Delta\{1, \ldots, K\}$ is the set of all probability distributions over the indexes of the arms). The sequence (φ_t) is referred to as an allocation strategy. 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 $X_{I_t,N_{I_t,t}}$ with the notation above. The primary task is to output at the end of each round t a policy $\psi_t \in \Delta\{1, \ldots, K\}$ to be played in a new one-shot instance if the environment sends some stopping signal meaning that the exploration phase is over. The information

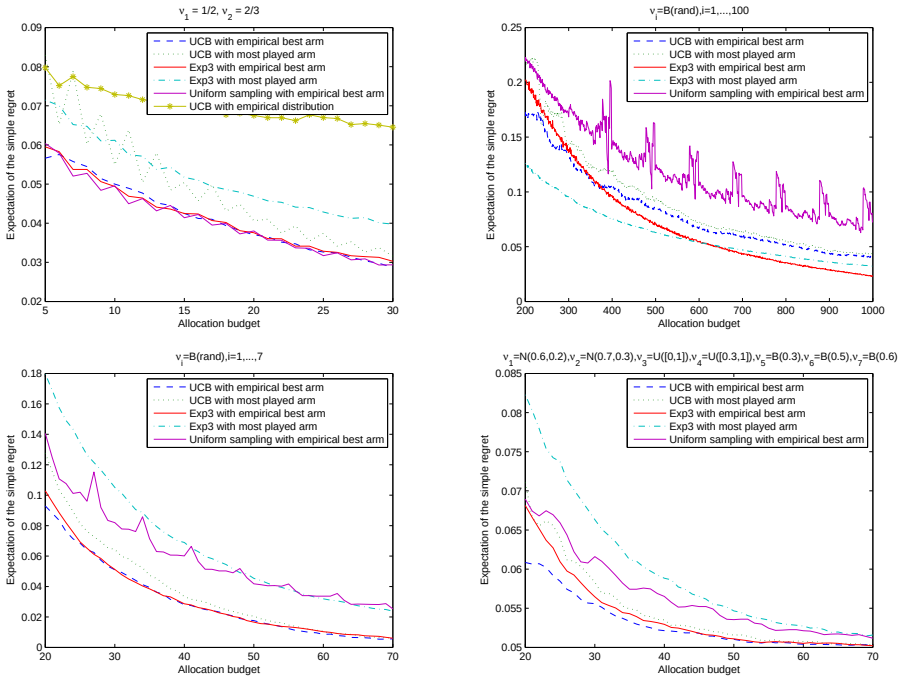

Figure 1: Experimental comparison of the exploration strategies introduced below.

available to the forecaster for choosing φ_t, respectively ψ_t, is formed by the $X_{j,s}$ for $j = 1, \ldots, K$ and $s = 1, \ldots, N_{j,t-1}$, respectively, $s = 1, \ldots, N_{j,t}$.

As we are only interested in the performances of the sequence (ψ_t) of policies, we call this problem the pure exploration problem for multi-armed bandits. The simple regret at round n of the policy $\psi_n = (\psi_{j,n})_{j=1,\ldots,N}$ is defined by $r_n = \mu^* - \sum_{j=1}^K \psi_{j,n}\mu_j$ where μ^* denote the expected reward of a best arm j^*. The simple regret is thus the expected regret on a new-one shot instance conditionally to the exploration phase.

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., Lai and Robbins (1985) or Auer et al. (2002a), and even $R_n = o(n)$ a.s., as follows, e.g., from Auer et al. (2002b, Theorem 6.3) together with a martingale argument. The cumulative regret is the sum of the instantaneous regrets $r'_t = \mu^* - \mu_{I_t}$, but the latter can be hardly related to the simple regrets r_t.

Goal: In this paper, we study the links between simple and cumulative regrets and show that, surprisingly enough and perhaps counter-intuitively, the strategies that are best in practice rely on the exploration–exploitation dilemma, whereas their assessment criterion, the simple regret, is only a matter of efficient exploration and involves no exploitation.

3 Simulation study

We start by indicating some surprising experimental results, which motivated the present work. We considered three different allocation strategies, uniform sampling (pull all arms, one after the other), EXP3 (see Auer et al., 2002b), and UCB1 (see Auer et al., 2002a) and three associated policies, empirical distribution, empirical best arm, and most played arm. We recall that EXP3 and UCB1 perform an exploration–exploitation tradeoff, while uniform sampling focuses on (uniform) exploration of the arms. To a given allocation, we may associate the policy that either is the empirical distribution of the arms sampled in the exploration phase, or (the Dirac mass on) the arm with best empirical mean at the end of the exploration phase, or (the Dirac mass on) the arm played most often in the exploration phase. These policies are described in detail respectively in Lemma 3, Section 5.2, and Lemma 4.

The resulting simple regrets are computed over $10\,000$ runs of each K–tuple of distributions and we plot their averages on Figures 1, which approximate well the expectations $\mathbb{E}r_n$. The distributions used for the simulations are given by Bernoulli distributions, uniform distributions, or Gaussian distributions (which are almost finitely supported), in number and with parameters depending on the experiment (see the captions of the different figures for a description of each K–tuple). In this extended abstract, we only offer a limited number of graphical illustrations of the performances, but mention that the situations illustrated below are typical.

Only Figure 1 top–left shows one empirical distribution policy, based on UCB1; the one based on EXP3 performs similarly. They are always worse than associated empirical best arm or empirical most played arm policies and this is why we do not report their performances on other pictures. Empirical distributions will thus be of theoretical interest only; they probably suffer in practice from being too conservative.

The ranking of the different strategies strongly depends on the number of arms. For $K = 2$ arms, Figure 1 top–left shows that the empirical best arm policy is the best one, and that its performances are almost independent of the underlying exploration strategy (uniform, UCB1, or EXP3). For small values of K (say, K between 3 and 10), Figure 1 bottom–left and bottom–right indicate that the best strategies are the ones that pick the empirical best arm after exploring with EXP3 or UCB1; UCB1 combined with the selection of the most played arm in the exploration phase is also an interesting competitor. When a large number of arms is available, EXP3 becomes the unique best exploration strategy, and the optimal associated policy is the most played arm for a small number of rounds and the empirical best arm for a larger number of rounds, as Figure 1 top–right reveals. This is maybe a surprising fact, since EXP3 is not designed for a stochastic, but an adversarial, environment.

In total, maintaining some exploration–exploitation trade-off even in the exploration phase seems worthwhile. A heuristic explanation would be that uniform sampling gives the same attention to all arms whereas forecasters designed to minimize the cumulative regret tend to focus on a much smaller sub-sample of arms, playing almost only the ones that are likely to be optimal.

Goal (continued) and structure of the paper: We aim at giving some (partial but more mathematical) explanations of these surprising facts in the rest of the paper. We first account for the intuition that it should not be the case that strategies trading off between exploration and exploitation can be efficient in such a full exploration problem (Section 4). We do so by studying distribution-dependent bounds. We then are able to qualify this statement by indicating some distribution-free bounds (Section 5).

4 Too small the cumulative regret is bad for the simple regret

Lemma 3 states in the next section that $\mathbb{E}r_n = \mathbb{E}R_n/n$ for the empirical distribution policy, and 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 bound on $\mathbb{E}R_n$, the worst the bound on $\mathbb{E}r_n$, no matter what the policies (ψ_n) are. This is interpreted as a consequence of the classical trade-off between exploration and exploitation. The design of (ψ_n) relies on an efficient exploration only, whereas the minimization of $\mathbb{E}R_n$ requires exploitation of the results of exploration considered as a side-task.

Theorem 1 (Main theorem) *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 $C \geq 0$ with $\mathbb{E}R_n \leq C\varepsilon(n)$, the simple regret of any policy (ψ_n) based on the allocation (φ_t) is such that for all sets of $K \geq 3$ (distinct, Bernoulli) distributions on the rewards, there exist a constant $D \geq 0$ with*

$$\mathbb{E}r_n \geq \frac{1}{2}\left(\min_{j:\Delta_j>0} \Delta_j\right) e^{-D\varepsilon(n)}$$

(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).

To get the point of this result, one should keep in mind that the typical rate of growth of the cumulative regrets of good algorithms, e.g. UCB1 of Auer et al. (2002a), is $\varepsilon(n) = \ln n$. This, as asserted in Lai and Robbins (1985), is the optimal rate. The policies based on such allocation strategies are bound to suffer a simple regret that decreases at best polynomially fast. For instance, it follows from Kocsis and Szepesvari (2006, Theorem 5) that the simple regret of the empirical best arm policy based on a UCB1 allocation decreases at a polynomial rate, and this is no accident. On the contrary, the empirical best arm policy based on a uniform exploration has a simple regret decreasing exponentially fast, as shown by Theorem 7. In addition, it follows from the theorem above and the trivial inequality $\mathbb{E}R_n \leq n$ that this latter exponential decrease is the best achievable rate for the simple regret.

Proof: The basic idea of the proof 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. 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 where those forecasters choose the worst arm instead of the best one. We consider now a set of $K \geq 3$ (distinct) Bernoulli distributions; actually, we only use below that their parameters are (up to a first relabeling) such that $\mu_1 > \mu_2 \geq \mu_3 \geq \ldots \geq \mu_K$ and $\mu_2 > \mu_K$, and thus, $\mu_2 > 0$.

Another layer of notation is needed. Fix a permutation σ of $\{1, \ldots, K\}$. For $i = 1$ (respectively, $i = K$), we denote by $\mathbb{P}_{i,\sigma}$ and $\mathbb{E}_{i,\sigma}$ the probability and expectation with respect to the K-tuple formed by the $\nu_{\sigma^{-1}(j)}$, where we replaced the best of them, indexed by $\sigma(1)$, by a Dirac measure on 0 (respectively, the best and worst of them, indexed by $\sigma(1)$ and $\sigma(K)$, by Dirac measures on 0). We provide a proof in five steps.

Step 1 lower bounds the maximum by an average,

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

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

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

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

$$\mathbb{E}_\sigma\left[1 - \psi_{\sigma(1),n}\right] \geq \mathbb{E}_{K,\sigma}\left[\left(1 - \psi_{\sigma(1),n}\right)\right] \ \mathbb{P}_{1,\sigma}\left\{\widehat{\mu}_{\sigma(K),n} = 0\right\} \ \mathbb{P}_\sigma\left\{\widehat{\mu}_{\sigma(1),n} = 0\right\} \ . \qquad (1)$$

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

$$\mathbb{P}_{1,\sigma}\left(\widehat{\mu}_{\sigma(K),n}=0\right)=\mathbb{E}_{1,\sigma}\left[(1-\mu_K)^{N_{\sigma(K),n}}\right]\geq(1-\mu_K)^{\mathbb{E}_{1,\sigma}N_{\sigma(K),n}}\ ,$$

where the equality can be seen by first 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 very definition of the latter), $(\mu_2-\mu_K)\,\mathbb{E}_{1,\sigma}N_{\sigma(K),n}\leq\mathbb{E}_{1,\sigma}R_n$; since by hypothesis, there exists a constant C such that for all σ, $\mathbb{E}_{1,\sigma}R_n\leq C\,\psi(n)$, we finally get

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

Step 4 proves that the third term in the right-hand side of (1) is more than

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

We denote by $W_n=(I_1,X_{I_1,1},\ldots,I_n,X_{I_n,N_{I_n,n}})$ the history up to time n. What follows is reminiscent of the techniques used in Mannor and Tsitsiklis (2004). We are interested in realizations $w_n=(i_1,x_{i_1,1},\ldots,i_n,x_{i_n,n_{i_n,n}})$ of the history such that whenever $\sigma(1)$ was played, it got a null reward. (We denote above by $n_{j,t}$ is the realization of $N_{j,t}$ corresponding to w_n, for all j and t.) The likelihood of such a w_n under $\mathbb{P}_\sigma$ is $(1-\mu_1)^{n_{\sigma(1),n}}$ times the one under $\mathbb{P}_{1,\sigma}$. Thus,

$$\mathbb{P}_{\sigma}\big\{\widehat{\mu}_{\sigma(1),n}=0\big\}=\sum\mathbb{P}_\sigma\left(W_n=w_n\right)=\sum(1-\mu_1)^{n_{\sigma(1),n}}\,\mathbb{P}_{1,\sigma}\left(W_n=w_n\right)=\mathbb{E}_{1,\sigma}\left[(1-\mu_1)^{N_{\sigma(1),n}}\right]$$

where the sums are over those histories w_n such that $x_{\sigma(1),t}=0$ for all $t=1,\ldots,n_{\sigma(1),n}$. The argument is concluded as before, first by Jensen's inequality and then, by using that $\mu_2\mathbb{E}_{1,\sigma}N_{\sigma(1),n}\leq\mathbb{E}_{1,\sigma}R_n\leq C\,\varepsilon(n)$ by definition of the regret and the hypothesis put on its control.

Step 5 concludes the proof by resorting to a symmetry argument to show that as far as the first terms of the right-hand side of (1) are concerned,

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

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 is 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 is such that $\sigma(1)$ equals the other such element. Formally,

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

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

A paradox? At this point there seems to be a contradiction between the experimental observations and the theory, since according to the rates of convergence in n, the simple regrets of the policies based on uniform allocation (decreasing exponentially fast, see Theorem 7) should be below the ones based on EXP3 or UCB1 allocations (that can be decreasing at best polynomially fast, as Theorem 1 indicates). But the distribution-dependent multiplicative constants play a role: in practice we observed this ranking only for simple regrets smaller than 10^{-10}, a precision for which little can be guaranteed in terms of correct numerical computations. Thus we believe that there are two regimes, a first one for small numbers of rounds n (this is the one observed in the simulations) and a second one for very large numbers of rounds. In the next section, distribution-free bounds turn out to be a good way to capture the good behavior of the simple regrets of UCB1 and EXP3 based strategies in the small–n regime.

5 Consideration of distribution-free bounds solve (partially) the paradox

Theorem 1 shows in particular that as long as distribution-dependent bounds are considered, no faster than exponential rates of decrease can be achieved for simple regrets. For distribution-free bounds, the rate worsens to $1/\sqrt{n}$. We start by indicating a general lower bound (a simple variation on the proof provided in Auer et al., 2002b, Appendix A), and then present some distribution-free upper bounds on the simple regret, which are almost optimal in the sense that they match the order of magnitudes of the lower bound up to log factors. While it was expected that uniform sampling associated to the empirical best arm policy was almost optimal, it is surprising that allocations with EXP3 or UCB1 can be almost optimal as well, whereas they are designed to minimize the cumulative regret.

Proposition 2 (simple variation on Auer et al., 2002b) *For all $n \geq 1$ and $K \geq 2$ such that $n > K/(4\ln(4/3))$, the simple regrets of any allocation strategy and any policy based on this allocation are bounded in a minimax sense as*

$$\inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \geq \frac{1}{32\sqrt{\ln(4/3)}}\sqrt{\frac{K}{n}}$$

where the infimum is taken over all (randomized) allocation strategies and all associated policies and the supremum over all K-tuples of probability distributions with support in $[0,1]$.

5.1 Bounds on the simple regrets of UCB1 and EXP3

Some of the bounds can be obtained in an automatic way from the bounds on the cumulative regrets via the following two lemmas. Only the proof of the second one deserves a word; it uses that if J is the random index of the most played arm (ties broken in some way), then $\Delta_J\, n/K \leq \Delta_J\, N_{J,n} \leq \sum_j \Delta_j\, N_{j,n} = R_n$ and the simple regret is $\mathbb{E}r_n = \mathbb{E}\Delta_J$. Our current bounds for the empirical best arm policy based on UCB1 or EXP3 allocations rely on concentration-of-the-measure methods and do not reach the $1/\sqrt{n}$ rate yet; for this reason, we do not report them here.

Lemma 3 *For all allocation strategies (φ_n), the sequence of policies (ψ_n), called the empirical distribution policies and defined, for all $n = 1, 2, \ldots$, by $\psi_n = (1/n)\sum_{t=1}^n \delta_{I_t}$ (where δ_j denotes the Dirac mass on arm j), is such that for all n, its simple regret satisfies $r_n = R_n/n$.*

Lemma 4 *For all allocation strategies (φ_n), the sequence of policies (ψ_n), called the empirical most played arm, and defined, for all $n = 1, 2, \ldots$, by $\psi_n = \delta_J$ where $J \in \operatorname{argmax}_j N_{j,n}$, is such that for all n, its simple regret satisfies $\mathbb{E}r_n \leq K\,\mathbb{E}R_n/n$.*

Corollary 5 *The simple regrets of the allocation strategies EXP3 of Auer et al. (2002b) and UCB1 of Auer et al. (2002a), combined with the empirical distribution policies, are respectively bounded by*

$$\inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leq 4\sqrt{\frac{K\ln K}{n}} \qquad \textit{and} \qquad \inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leq \sqrt{\frac{K\,(8\ln n + 1 + \pi^2/3)}{n}}\,.$$

Proof: The bounds follows from the distribution-free bounds on the cumulative regrets via Lemma 3. We provide here such a bound for UCB1, the one for EXP3 being given in Auer et al. (2002b). It can be extracted from the proof of Auer et al. (2002a, Theorem 1) that for all suboptimal arm j,

$$\mathbb{E}N_{j,n} \leq \frac{8\ln n}{\Delta_j^2} + 1 + \frac{\pi^2}{3} \qquad \text{hence} \qquad \mathbb{E}R_n = \sum_{j:\Delta_j>0} \Delta_j\,\mathbb{E}N_{j,n} \leq \sqrt{8\ln n + 1 + \frac{\pi^2}{3}} \sum_{j:\Delta_j>0} \sqrt{\mathbb{E}N_{j,n}}\,.$$

The conclusion follows by the concavity of the square root, which entails $\sum \sqrt{\mathbb{E}N_{j,n}} \leq \sqrt{Kn}$. ■

Corollary 6 *The simple regrets of the allocation strategies* EXP3 *of Auer et al. (2002b) and* UCB1 *of Auer et al. (2002a), combined with the policy given by the choice of the empirical most played arm, are respectively bounded by*

$$\inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leq 4K\sqrt{\frac{K\ln K}{n}} \qquad \textit{and} \qquad \inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leq K\sqrt{\frac{(8\ln n + 1 + \pi^2/3)}{n}}\,.$$

Proof: The first bound is obtained via Lemma 4. For the second one, a sharper argument uses, as in the proof of Lemma 4, that $\mathbb{E}N_{j,n} \geq \mathbb{P}\{J=j\}\, n/K$ and the same upper bound on $\mathbb{E}N_{j,n}$ as above to get

$$\Delta_j\sqrt{\mathbb{P}\{J=j\}} \leq \sqrt{\frac{K\,(8\ln n + 1 + \pi^2/3)}{n}}\,;$$

the proof is concluded by concavity again. ■

5.2 Bounds on the simple regret of uniform sampling

Formally, uniform sampling consists in choosing the allocations $\varphi_t = \delta_{[t \bmod K]}$ where $[t \bmod K]$ denotes the value of t modulo K. Thus, arm j is played at rounds $j,\ j+K,\ j+2K\ \ldots$. We now denote, for $n \geq K$ and $j = 1,\ldots,K$,

$$\widehat{\mu}_{j,n} = \frac{1}{\lfloor n/K \rfloor}\sum_{s=1}^{\lfloor n/K \rfloor} X_{j,s}$$

the mean reward of j on the first $K\lfloor n/K \rfloor$ rounds. ($\lfloor n/K \rfloor$ denotes the lower integer part of n/K. We discard here some final rounds for all arms to have been played equally often whenever a new decision is made.) The associated empirical best arm policy is defined by $\psi_1 = \ldots = \psi_{K-1}$ equal to the uniform distribution and $\psi_n = \delta_{j_n^*}$ where $j_n^* \in \operatorname{argmax}_{j=1,\ldots,N} \widehat{\mu}_{j,n}$ for $n \geq K$ (ties broken in some way). The proof of the following theorem can be found in the appendix.

Theorem 7 *The uniform sampling allocation associated to the empirical best arm policy ensures that simple regrets are bounded, respectively in a distribution-dependent and in a distribution-free sense, by*

$$\mathbb{E}r_n \leq \sum_{j:\Delta_j>0} \Delta_j\, e^{-\Delta_j^2\lfloor n/K \rfloor/2} \qquad \textit{and} \qquad \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leq 2\sqrt{\frac{2K\ln K}{n}}\,.$$

References

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

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, 2002b.

P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. Technical Report Inria-00136198, INRIA, 2007.

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.

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

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

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

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

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.

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

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