Bandit Exploration

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

Sébastien Bubeck is a student author and wishes to be considered for the Mark Fulk award.

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 not in terms of their cumulative regrets, as is usually the case, but through quantities referred to as simple regrets. The latter are related to the (expected) gains of the decisions that the strategies would recommend for a new one-shot instance of the same multi-armed bandit problem. Here, exploration is only constrained by the number of available rounds (not necessarily known in advance), in contrast to the case when cumulative regrets are considered and when exploitation needs to be performed at the same time. We start by indicating the links between simple and cumulative regrets. A small cumulative regret entails a small simple regret but too small a cumulative regret prevents the simple regret from decreasing exponentially towards zero, its optimal distribution-dependent rate. We therefore introduce specific strategies, for which we prove both distribution-dependent and distribution-free bounds. A concluding experimental study puts these theoretical bounds in perspective and shows the interest of nonuniform exploration of the arms.

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, a sequential decision problem where, at each stage, the forecaster has to pull one of K given arms and gets a reward. We consider here a stochastic version in which each arm is parameterized by a probability distribution and provides, when pulled, a reward drawn at random according to this distribution. The problem was first considered by Robbins (1952), who derives strategies that asymptotically attain a per-round reward equal to the expected reward of the best arm. See Berry and Fristedt (1985) for a review of the classical (statistical) results. The problem turned out to be fundamental as well in artificial intelligence and machine learning but it was actually originally motivated by medical statistics, see Schlag (2006) for more historical details and references to the rich corresponding literature. The typical case is the test of a new treatment against the best treatment known so far. Of course, some exploration is needed, so that the test samples for each treatment are large enough; this is the only way to guarantee a correct estimation of their respective efficiencies. But since human beings are involved, the cost of picking the wrong treatment is high (the associated reward would equal a large negative value); therefore, exploration is potentially risky and exploitation has to be done most of the time. This is captured by a quantity called the cumulative regret, defined as the difference between the per-round reward of the best arm (treatment, in the medical case) and the one of the considered strategy. Among others, the UCB strategies (standing for "upper confidence bounds") of Auer et al. (2002a) perform a good exploration–exploitation tradeoff in terms of this cumulative regret.

When exploration involves costs not measured in terms of rewards but rather in terms of resources (e.g., memory or CPU), the performances of a strategy have to be assessed in a different way. As illustrated below, it is then natural that the forecaster allocates sequentially its resources to explore the arms to his convenience and outputs the index of an arm when these resources have been all used. Finally, the recommended arm is used on a new one-shot instance of the same bandit problem. Therefore, a strategy is assessed here in terms of the performances of the recommended arm on the 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). This pure exploration problem was referred to as "budgeted multi-armed bandit problem" in the COLT'04 open problem by Madani et al. (2004). They adress it from a Bayesian and Markov decision processes point of view, discuss some natural strategies, and point out difficulties.

A concrete example for such a pure exploration problem

*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.

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 (standing for "UCB for trees") of Kocsis and Szepesvari (2006) and the BAST strategy (standing for "bandits algorithms for tree search") 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 focus on exploration and minimizing the simple regret are expected to be the stone for an improvement of these results. Another valuable example is provided by the setting of so-called computer experiments (see, for instance, Santner et al., 2003 for a review), e.g., the optimization of some complex function that can only be evaluated through a (possibly noisy) black-box at a computational only fixed cost.

Here, the number n of rounds available to the forecaster is fixed in advance but may be unknown to the forecaster. It is determined in the previous examples by the time given to the forecaster with respect to the computational power of the computer he is using. Of related interest is the setting of Even-Dar et al. (2002) and Mannor and Tsitsiklis (2004), in which the forecaster can perform exploration during a random number of rounds T (which he chooses) and aims 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. Performances are therefore not measured by some regret in this case.

Schlag (2006) solves the pure exploration problem for the case of two arms and rewards given by probability distributions over $[0,1]$. He shows that the minimax value of the simple regret (where the supremum is taken with respect to all possible pairs of probability distributions and the infimum is over all possible strategies of the forecaster) is achieved by a rule called binomial average rule, sampling the two arms equally often and basically recommending the one with higher empirical mean. He also computes this minimax value as a function of n and proves that it converges to zero at a $(1/\sqrt{n})$–rate. The article is concluded by pointing out some difficulties for the case of three or more arms.

Static allocations of resources (i.e., the use of a strategy determined before the samplings) were considered in Schlag (2006) for the case of two probability distributions over compact intervals and in Chen et al. (2000) for an arbitrary number of Gaussian distributions. In the latter case, arms are optimally sampled (from the static viewpoint) depending on their (known) variances.

Structure of the paper: We present formally the model in Section 2 and define cumulative and simple regrets. Section 3 investigates the links between these regrets and indicates that upper bounds on the cumulative regrets lead not only to upper bounds on the simple regrets (Section 3.1), but also to lower bounds on them. Actually, Section 3.2 states the main result of the paper; strategies with too good (e.g., logarithmically only increasing) distribution-dependent bounds on their cumulative regrets lead to strategies with simple regrets decreasing at best at a polynomial rate. This is in contrast with the results of Section 4, which introduces specific forecasters and shows an exponentially fast decrease of the simple regrets of some of these strategies. For the cumulative regrets to be minimized, exploitation is needed and the price to pay for that is a worse control on the simple regrets, as captured by Theorem 4. Section 5 qualifies however some of the gaps in the orders of magnitude of distribution-dependent bounds exhibited in the previous sections and explains that, as far as distribution-free bounds are considered, most of the previous forecasters have simple regrets decreasing at the optimal $1/\sqrt{n}$–rate (up to logarithmic factors). This is similar to what can be said for distribution-dependent versus distribution-free bounds for the cumulative regret. (By passing, we also exhibit distribution-free bounds for the UCB1 strategy of Auer et al., 2002a.) The simulation study of Section 6 puts these theoretical results in perspective and shows the interest of non-uniform exploration of the arms.

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 $n \geq 1$, we denote by $N_{j,n}$ the number of times j was pulled from rounds 1 to n, and by $X_{j,1}, X_{j,2}, \ldots, X_{j,N_{j,n}}$ 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 $\Phi_t \in \Delta\{1, \ldots, K\}$ to be played in a new one-shot instance if the environment sends some stopping signal meaning that the exploration phase is over.

As we are only interested in the performances of the sequence (Φ_n) of policies, we call this problem the pure exploration problem for multi-armed bandits. Figure 1 summarizes the description of the sequential game and points out that the information available to the forecaster for choosing φ_t, respectively Φ_t, is formed by the $X_{j,s}$ for $j = 1, \ldots, K$ and $s = 1, \ldots, N_{j,t-1}$, respectively, $s = 1, \ldots, N_{j,t}$. Formally, we define the simple regret at round n (of the policy

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

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

(3) the forecaster outputs a policy $\Phi_t \in \Delta\{1, \ldots, K\}$;

(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

Φ_n) by

$$r_n = r\big(\Phi_n\big) = \mu^* - \mu_{\Phi_n}$$

where $\mu^* = \mu_{j^*} = \max_{j=1,\ldots,K} \mu_j$ and $\mu_{\Phi_n} = \sum_{j=1,\ldots,K} \Phi_{j,n}\, \mu_j$

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 policy $\Phi_n = (\Phi_{j,n})_{j=1,\ldots,N}$.

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 quantities $r'_t = \mu^* - \mu_{I_t}$ are sometimes called instantaneous regrets. They differ from simple regrets r_t and in particular, $R_n = r'_1 + \ldots + r'_n$ is in general not equal to $r_1 + \ldots + r_n$. Lemma 1 and Theorem 4 will however indicate some connections between r_n and R_n.

Goal: We focus here on simple regrets r_n and ask for strategies ensuring that $\mathbb{E}r_n = o(1)$.

3 Links between cumulative and simple regrets

In this section, we show how R_n and r_n are related. We first state the straightforward upper bound $\mathbb{E}r_n \leq \mathbb{E}R_n/n$, which holds for suitable choices of (Φ_n) and shows that it is indeed possible to guarantee $\mathbb{E}r_n = o(1)$.

More interestingly, we then prove that upper bounds on $\mathbb{E}R_n$ lead to lower bounds on $\mathbb{E}r_n$; the better the guaranteed bound on $\mathbb{E}R_n$, the worst the bound on $\mathbb{E}r_n$. This is interpreted as a consequence of the classical trade-off between exploration and exploitation. The design of (Φ_n) relies on an efficient exploration only, whereas the minimization of $\mathbb{E}R_n$ requires exploitation of the results of the exploration considered as a side-task.

3.1 A small cumulative regret entails a small simple regret...

Classical strategies for multi-armed bandits minimize the cumulative regret. We consider below the UCB family of strategies introduced in Auer et al. (2002a) for a stochastic environment and the EXP3 family suited for an adversarial environment, see Auer et al. (2002b). Other strategies have been studied like, e.g., the GREEN strategy of Allenberg et al. (2006), but for simplicity we focus on the two families mentioned above.

These strategies consist only in a sequence of allocation strategies (φ_t). We construct the associated sequences of policies (Φ_n) and $(\overline{\Phi}_n)$ as either the sequence of the empirical distributions of the I_t or the sequence of moving averages of the φ_t; for all $n = 1, 2, \ldots,$

$$\Phi_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{I_t} \quad \text{and} \quad \overline{\Phi}_n = \frac{1}{n}\sum_{t=1}^{n} \varphi_t \tag{1}$$

where δ_j denotes the Dirac mass on arm j.

The next lemma is almost a triviality and follows from the linearity of r_n in the policy Φ_n and from the rewriting of $\mathbb{E}R_n$ as

$$\mathbb{E}R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right] = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{j=1}^{K} \varphi_{j,t}\mu_j\right]$$

(as can be seen by taking expectations of conditional expectations).

Lemma 1 *For all allocation strategies (φ_t), the sequences (Φ_n) and $(\overline{\Phi}_n)$ of policies obtained by taking the empirical distributions and the moving averages (1) are such that for all $n = 1, 2, \ldots,$*

$$r_n\big(\Phi_n\big) = \frac{R_n}{n} \quad \textit{and} \quad \mathbb{E}\left[r_n\big(\overline{\Phi}_n\big)\right] = \frac{\mathbb{E}R_n}{n}\,.$$

In particular, for both sequences of policies, if $\mathbb{E}R_n = o(n)$, then $\mathbb{E}r_n = o(1)$.

For further reference, we illustrate this lemma with the strategies UCB1 of Auer et al. (2002a) and a simple modification of EXP3 (without the mixing step). There is a fundamental difference between these two types of bounds, that will be further illustrated in Sections 4 and 5. The bound for UCB1 depends on the distributions $\nu_1, \ldots, \nu_K$ of the arms whereas the one of EXP3 holds for all possible distributions over a given interval, say $[0, 1]$ for simplicity.

For the sake of completeness, we recall the statement of UCB1. In the first K rounds, each arm is played once, that is, $\varphi_t = \delta_t$ is the Dirac mass on t for $t = 1, \ldots, K$. Then, for $t \geq K$ and $j = 1, \ldots, K$, we denote by

$$N_{j,t} = \sum_{s=1}^{t} \mathbb{I}_{\{I_s = j\}} \quad \text{and} \quad \widehat{\mu}_{j,t} = \frac{1}{N_{j,t}}\sum_{s=1}^{N_{j,t}} X_{j,s}$$

the number of rounds arm j was pulled before round t and the mean reward of j on these rounds. Now, $\varphi_{t+1} = \delta_{j_t^*}$, where

$$j_t^* \in \operatorname*{argmax}_{j=1,\ldots,K} \widehat{\mu}_{j,t} + \sqrt{\frac{2\ln t}{N_{j,t}}} \tag{2}$$

(ties broken by choosing, for instance, the arm with smallest index). UCB1 has a deterministic allocation strategy, the associated averages $(\Phi_n) = (\overline{\Phi}_n)$ are however given by non-degenerated probability distributions. Auer et al. (2002a, Theorem 1) together with Lemma 1 (and the bound $\Delta_j \leq 1$) implies the following. We denote in the sequel $\Delta_j = \mu^* - \mu_j$ the gap between the expected reward of the best arm and the one of arm j.

Theorem 2 (Auer et al., 2002a) *For all probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$, UCB1 ensures that its expected regret is bounded as*

$$\mathbb{E}R_n \leq 8\left(\sum_{j:\mu_j<\mu^*} \frac{1}{\Delta_j}\right)\ln n + \left(1+\frac{\pi^2}{3}\right)\left(\sum_{j=1,\ldots,K} \Delta_j\right)$$

for all $n \geq 1$. Thus the expected simple regrets of the sequence of empirical distributions and moving averages (1) are bounded as

$$\mathbb{E}r_n \leq 8\left(\sum_{j:\mu_j<\mu^*} \frac{1}{\Delta_j}\right)\frac{\ln n}{n} + \left(1+\frac{\pi^2}{3}\right)\frac{K}{n}$$

for all $n \geq 1$.

We state this result mostly to point out in Section 4 that the rate $(\ln n)/n$ is not the optimal order of magnitude of the expected simple regret as far as distribution-dependent bounds are considered, and that specific algorithms need therefore to be constructed.

We now turn to EXP3-type forecasters. It has been noted since Auer et al. (2002b) that to obtain bounds on the (expectation of the) cumulative regret, EXP3 does not need the mixing step. We describe the version recalled, e.g., in the introduction of Stoltz (2005), see also Juditsky et al. (2008). φ_1 is the uniform distribution and for $t \geq 2$, we define φ_t component-wise as

$$\varphi_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{j,s}\right)}{\sum_{k=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{k,s}\right)}$$

for all $j = 1, \ldots, K$, where

$$\eta_t = \sqrt{\frac{2\ln K}{K\,t}} \quad \text{and} \quad \widehat{\ell}_{k,s} = \frac{1 - X_{k,N_{k,s}}}{\varphi_{k,s}} \mathbb{I}_{\{I_s=k\}}$$

are the estimated losses associated to the rewards. The bounds of Stoltz (2005, Theorem 2.7) and Lemma 1 imply the following.

Theorem 3 (variation of Auer et al., 2002b) *For all probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$, the variant of EXP3 recalled above ensures that its expected regret is bounded as*

$$\mathbb{E}R_n \leq \sqrt{8(n+1)K\ln K}$$

for all $n \geq 1$. Thus the expected simple regrets of either of the sequences of empirical distributions and moving averages (1) are bounded as

$$\mathbb{E}r_n \leq 4\sqrt{\frac{K\ln K}{n}}$$

for all $n \geq 1$.

3.2 ... but too small a cumulative regret forces too large a simple regret

The main result in this section indicates that the better the upper bound on the cumulative regret of a strategy, the larger the lower bound on its simple regret. Like in Mannor and Tsitsiklis (2004), since we are interested in lower bounds in this section, we mostly consider Bernoulli distributions. The function ψ of interest has to be thought of as the distribution-dependent upper bound on the order of magnitude of the regret, e.g., $\psi(n) = \ln n$ in Theorem 2.

Theorem 4 (Main theorem) *For all allocation strategies (φ_t) and all functions $\psi : \{1,2,\ldots\} \to \mathbb{R}$ such that*

> *for all (Bernoulli) distributions $\nu_1, \ldots, \nu_K$ on the rewards, there exists a constant $C \in \mathbb{R}^+$ with $\mathbb{E}R_n \leq C\psi(n)$,*

the simple regret of any policy (Φ_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*
>
> $$\mathbb{E}r_n \geq \frac{1}{2}\left(\min_{j:\Delta_j>0} \Delta_j\right) e^{-D\psi(n)}$$
>
> *(up to a relabelling $\nu_1, \ldots, \nu_K$ of the considered distributions into $\nu_{\pi(1)}, \ldots, \nu_{\pi(K)}$ for some permutation π).*

In particular, the polynomial decrease of the simple regret in Theorem 2 is not an accident, our main theorem shows that this needs to be the case in view of the good (logarithmic bound) performances of UCB1 in terms of the cumulative regret $\mathbb{E}R_n$. This is why a specific strategy is constructed and studied in Section 4, with exponential convergence rate to 0. Its allocation strategy relies on a heavy exploration and no exploitation. For the cumulative regret $\mathbb{E}R_n$ to be minimized, exploitation is needed and the price to pay for that is, at least from a theoretical viewpoint, a worse control on the simple regret, as captured by Theorem 4. (This statement is however qualified in practice for large numbers of rounds n, in Section 6.) This may be worth noticing in all applications where controlling the cumulative regret is crucial (one can think of the evaluation of new medical treatments) and thus no efficient exploration can be performed.

As a warm-up and to illustrate some of the techniques needed in the proof of Theorem 4, we start with two distribution-dependent lower bounds on the simple regret, first, one for specific policies and second, one for the general case of all policies. The bound of Theorem 6 is of independent interest, since it shows that the distribution-dependent rate $(\ln n)/n$ of Theorem 2 is suboptimal. Only then we prove Theorem 4.

For a given $n \geq 1$, we say that a policy Φ_n never plays an arm with zero empirical mean whenever it puts no probability mass ($\Phi_{j,n} = 0$) on arms j with empirical means $\widehat{\mu}_{j,n} = 0$, provided there is at least another arm with non-zero empirical mean. (In case all arms have zero empirical mean, then we impose, for the time being, that Φ_n puts equal weights on all arms.)

Lemma 5 *For all $n \geq 1$ and all policies Φ_n that never play an arm with zero empirical mean, for all distributions $\nu_1, \ldots, \nu_K$ of the rewards such that there is a single best distribution ν^* given by a Bernoulli distribution, the expected simple regret is lower bounded by*

$$\mathbb{E}r_n \geq \frac{K-1}{K} \left(\min_{j:\Delta_j>0} \Delta_j \right) e^{n \ln(1-\mu^*)} .$$

Proof: The simple regret is lower bounded in terms of the probability of not choosing the (unique) optimal arm j^*, which is in turn bounded by the probability that the best arm has a zero empirical mean (up to a factor of $(K-1)/K$, that takes into account the case of a tie between all arms),

$$\begin{aligned}
\mathbb{E}r_n &= \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}\Phi_{j,n} \\
&\geq \left(\min_{j:\Delta_j>0} \Delta_j \right) \mathbb{E}\left[1 - \Phi_{j^*,n}\right] \\
&\geq \left(\min_{j:\Delta_j>0} \Delta_j \right) \mathbb{E}\left[\left(1 - \Phi_{j^*,n}\right) \mathbb{I}_{\{\widehat{\mu}_{j^*,n}=0\}} \right] \\
&\geq \frac{K-1}{K} \left(\min_{j:\Delta_j>0} \Delta_j \right) \mathbb{P}\left\{\widehat{\mu}_{j^*,n} = 0\right\}
\end{aligned}$$

where we used that when j^* has an average reward $\widehat{\mu}_{j^*,n} = 0$, it is not played with probability at least $1 - 1/K$ (it could only be played in the case when all arms would have zero empirical average rewards). Now, conditionally to the drawn actions $I_1, \ldots, I_n$, and using the fact that ν^* is a Bernoulli distribution,

$$\mathbb{P}\left\{\widehat{\mu}_{j^*,n} = 0 \;\middle|\; I_1, \ldots, I_n\right\} = (1-\mu^*)^{N_{j^*,n}} \geq (1-\mu^*)^n ,$$

which concludes the proof by integration with respect to the conditioning. ■

We now turn to the general case, for which we are not willing to put any restriction on the policies Φ_n. Note that distribution-dependent lower bounds suffer from the general drawback that we can never prevent naive strategies like "play always the first arm" to be efficient despite all against some particular K-tuples of distributions of the arms. Symmetry is a way to deal with that. We therefore consider sets of distributions over the arms with K elements and run the forecaster over all possible K-tuples obtained from this set. In the sequel, we thus fix a set of K distributions, $\{\nu_1, \ldots, \nu_K\}$, and, with no loss of generality, we index them so that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$. For all permutations σ over $\{1, \ldots, K\}$, we denote by $\mathbb{P}_\sigma$ and $\mathbb{E}_\sigma$ the probability and expectation when the distributions of the arms are given by the K-tuple $\nu_{\sigma^{-1}(1)}, \ldots, \nu_{\sigma^{-1}(K)}$. In this K-tuple, the best arm has index $\sigma(1)$, the second best is $\sigma(2)$, and so on. We are now ready to state our policy-independent lower bound.

Theorem 6 *For all $n \geq 1$ and policies Φ_n, for all sets of distributions of rewards given by $K \geq 3$ Bernoulli distributions with parameters $\mu_1 > \mu_2 \geq \mu_3 \ldots \geq \mu_K$, the expected simple regret is lower bounded (up to a permutation of the arms) by*

$$\max_\sigma \mathbb{E}_\sigma r_n \geq \frac{1}{2} \left(\mu_1 - \mu_2\right) e^{n(\ln(1-\mu_1)+\ln(1-\mu_K))}$$

where the maximum is taken with respect to all permutations σ over $\{1, \ldots, K\}$.

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 n) 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 instead of the best arm.

Another layer of notation will be used in the proof. It could still be avoided here, but will be necessary for the proof of Theorem 4. 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 first use that a maximum is larger than a mean and extract from the previous proof a lower bound on simple regret in terms of incorrect selection of a best arm,

$$\max_\sigma \mathbb{E}_\sigma r_n \geq \frac{1}{K!} \sum_\sigma \mathbb{E}_\sigma r_n \geq \frac{\Delta}{K!} \sum_\sigma \mathbb{E}_\sigma \left[1 - \Phi_{\sigma(1),n}\right] \quad (3)$$

where we denoted $\Delta = \min_{j:\Delta_j>0} \Delta_j = \mu_1 - \mu_2$. For all σ,

$$\begin{aligned}
&\mathbb{E}_\sigma \left[1 - \Phi_{\sigma(1),n}\right] \\
&\quad \geq \mathbb{E}_\sigma \left[\left(1 - \Phi_{\sigma(1),n}\right) \mathbb{I}_{\{\widehat{\mu}_{\sigma(1),n}=0\}} \right] \\
&\quad = \mathbb{E}_\sigma \left[\left(1 - \Phi_{\sigma(1),n}\right) \;\middle|\; \widehat{\mu}_{\sigma(1),n} = 0 \right] \times \mathbb{P}_\sigma \left\{\widehat{\mu}_{\sigma(1),n} = 0\right\} \\
&\quad = \mathbb{E}_{1,\sigma} \left[\left(1 - \Phi_{\sigma(1),n}\right) \right] \mathbb{P}_\sigma \left\{\widehat{\mu}_{\sigma(1),n} = 0\right\} \\
&\quad = \mathbb{E}_{1,\sigma} \left[\left(1 - \Phi_{\sigma(1),n}\right) \right] \mathbb{E}_\sigma \left[\left(1 - \mu_1\right)^{N_{\sigma(1),n}} \right] \\
&\quad \geq \mathbb{E}_{1,\sigma} \left[\left(1 - \Phi_{\sigma(1),n}\right) \right] \left(1 - \mu_1\right)^n ,
\end{aligned}$$

where we used for the third step the fact that $\mathbb{P}_{1,\sigma}$ is the same as $\mathbb{P}_\sigma$, except that it ensures that arm $\sigma(1)$ has zero reward throughout, and subsequent steps are similar to the end of the proof of Lemma 5. We can obviously iterate the argument, and get, by considering the worst arm,

$$\begin{aligned}
&\mathbb{E}_{1,\sigma} \left[1 - \Phi_{\sigma(1),n}\right] \\
&\geq \mathbb{E}_{1,\sigma} \left[\left(1 - \Phi_{\sigma(1),n}\right) \;\middle|\; \widehat{\mu}_{\sigma(K),n} = 0 \right] \times \mathbb{P}_{1,\sigma} \left\{\widehat{\mu}_{\sigma(K),n} = 0\right\} \\
&\geq \mathbb{E}_{K,\sigma} \left[\left(1 - \Phi_{\sigma(1),n}\right) \right] \left(1 - \mu_K\right)^n .
\end{aligned}$$

Putting things together, we have proved

$$\begin{aligned}
&\sum_\sigma \mathbb{E}_\sigma \left[1 - \Phi_{\sigma(1),n}\right] \\
&\quad \geq \left(1 - \mu_1\right)^n \left(1 - \mu_K\right)^n \sum_\sigma \mathbb{E}_{K,\sigma} \left[1 - \Phi_{\sigma(1),n}\right] .
\end{aligned}$$

The proof is concluded by showing that by symmetry

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

and substituting this result in (3).

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,

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

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

We are now ready to prove Theorem 4.

Proof: We consider here a set of $K \geq 3$ (distinct) Bernoulli distributions; actually, we only use below that their parameters are (up to a first relabelling) such that $\mu_1 > \mu_2 \geq \mu_3 \ldots \geq \mu_K$, $\mu_2 > \mu_K$, and thus, $\mu_2 > 0$. We start with the following inequality, extracted from the proof of Theorem 6,

$$\begin{aligned}
\max_{\sigma} \mathbb{E}_{\sigma} r_n \quad &\geq \quad \frac{\mu_1-\mu_2}{K!}\sum_{\sigma} \mathbb{E}_{K,\sigma}\left[1-\Phi_{\sigma(1),n}\right] \\
&\qquad \times \mathbb{P}_{\sigma}\left\{\widehat{\mu}_{\sigma(1),n}=0\right\}\mathbb{P}_{1,\sigma}\left\{\widehat{\mu}_{\sigma(K),n}=0\right\}\,.
\end{aligned}$$

The last probabilities are bounded, for each permutation σ, by

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

where the equality comes from the proof of Theorem 6 and 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),

$$(\mu_2-\mu_K)\,\mathbb{E}_{1,\sigma}N_{\sigma(K),n} \leq \mathbb{E}_{1,\sigma}R_n \leq C\,\psi(n)\,;$$

we used that by hypothesis, there exists a constant C such that for all σ, $\mathbb{E}_{1,\sigma}R_n \leq C\,\psi(n)$. Substituting this inequality, we get for all σ,

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

We show below that one also has

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

Putting things together and resorting to the same symmetry argument as at the end of the proof of Theorem 6, we then will have proved

$$\begin{aligned}
&\max_{\sigma} \mathbb{E}_{\sigma} r_n \\
&\quad \geq \quad \frac{\mu_1-\mu_2}{K!}\,(1-\mu_1)^{C\psi(n)/\mu_2}\,(1-\mu_K)^{C\psi(n)/(\mu_2-\mu_K)} \\
&\qquad \times \sum \mathbb{E}_{K,\sigma}\left[1-\Phi_{\sigma(1),n}\right] \\
&\quad \geq \quad \frac{\mu_1-\mu_2}{2}\,(1-\mu_1)^{C\psi(n)/\mu_2}\,(1-\mu_K)^{C\psi(n)/(\mu_2-\mu_K)}
\end{aligned}$$

which yields the claimed result. The proof is thus concluded by studying the term involving $\{\widehat{\mu}_{\sigma(1),n}=0\}$. 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 reminiscient of the techniques used in Mannor and Tsitsiklis (2004). We are insteresting 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,

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

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

$$\mu_2\mathbb{E}_{1,\sigma}N_{\sigma(1),n} \leq \mathbb{E}_{1,\sigma}R_n \leq C\,\psi(n)$$

by definition of the regret and the hypothesis put on its control. ■

4 Distribution-dependent bounds

4.1 Uniform sampling and empirical successes

The previous section shows that specific forecasters need to be constructed for the pure exploration problem for multi-armed bandits. We study first the simplest of them, given by uniform sampling for the allocation strategy and empirical successes for the associated policy. We show an exponential decrease of its simple regrets towards 0, which is the best possible rate in view of Theorem 6.

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, $j+K$, $j+2K$ .... 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 policy, called empirical successes, is defined by $\Phi_1 = \ldots = \Phi_{K-1}$ equal to the uniform distribution and

$$\Phi_n = \delta_{j_n^*} \quad \text{where} \quad j_n^* \in \operatorname*{argmax}_{j=1,\ldots,N} \widehat{\mu}_{j,n} \tag{4}$$

for $n \geq K$ (ties broken in some way). We propose two bounds, the first one is sharper in the case when there are few arms and the gaps Δ_i can take extremal values (i.e., at least one value close to 0 and another one close to 1). The second one is suited for large n.

Theorem 7 *The uniform sampling allocation associated to the empirical successes policy ensures that the simple regrets are bounded by*

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

for all $n \geq K$; and by

$$\mathbb{E}r_n \leq \left(\max_{j=1,\ldots,K} \Delta_j\right) \exp\left(-\frac{1}{8}\left\lfloor \frac{n}{K}\right\rfloor \min_{j:\Delta_j>0} \Delta_j^2\right)$$

for all

$$n \geq \left(1 + \frac{8\ln K}{\min_{j:\Delta_j>0}\ \Delta_j^2}\right) K \, .$$

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

$$\mathbb{E}r_n = \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}\Phi_{j,n} \leq \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \widehat{\mu}_{j^*,n}\big\}$$

where the upper bound follows from the fact that to be the best empirical arm, an arm j must have performed, in particular, better than the mean best arm j^*. We now apply Hoeffding's inequality (for i.i.d. random variables, see Hoeffding, 1963). $\widehat{\mu}_{j,n} - \widehat{\mu}_{j^*,n}$ 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

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

and the first result follows.

The second inequality is proved by resorting to a sharper concentration argument, namely, the method of bounded differences, see McDiarmid (1989), see also Devroye and Lugosi (2001, Chapter 2). The proof can be found in appendix. ■

4.2 Empirical successes policy for other allocation strategies

Because we noticed by preliminary simulations, reported in Section 6, that the previous uniform sampling was often not the strategy with best practical performances, we now study the performances of the empirical successes policy (4) when the allocation strategy (φ_t) is not given by uniform sampling. The theoretical bounds will however be worse than the one of Theorem 7, and this will be explained by the lower bound on the performances given by Theorem 4.

4.2.1 UCB1 as allocation strategy

Since the proof of the theorem of this section will resort to concentration inequalities, we need to ensure that all arms are sampled sufficiently often each. Note that the following lemma indicates a deterministic lower bound on the number of times each arm is played. We provide the statement (and proof, see the appendix) only for $K = 2$ arms, but we believe that it extends to the case of more arms. Actually, Kocsis and Szepesvari (2006, Theorem 3) states a similar result, the proof being omitted there for the sake of space, but it is unclear whether their bound is uniform in all distributions over the arms, as we will need for later purposes, in Section 5.1, to get distribution-free bounds.

Lemma 8 *In the case of $K = 2$ arms, for all pairs of distributions ν_1 and ν_2, UCB1 pulls each arm, during the first $n \geq 3$ rounds, at least*

$$T_n = 2\ln\left((n-1)\left(1 - \sqrt{\frac{2\ln(n-1)}{n-1}}\right)\right)$$

times; consequently, $T_n \geq \log_2 n$ for $n \geq 21$.

This lemma ensures in particular that the following theorem is of interest when UCB1 is the allocation strategy and $n \geq 21$. Again, a similar result is provided by Kocsis and Szepesvari (2006, Theorem 5); there, however, the leading constant is not explicitly computed and the proof is omitted, again, for the sake of space. The leading constant we propose below is not suited for the needs of Section 5.1 because of the (form of the) dependency of the constant in the parameters Δ. A refined analysis will be needed there. For the moment, we point out that the following theorem illustrates the lower bound proposed by Theorem 4, in view of Theorem 2. We knew in advance that no faster than a polynomial distribution-dependent rate could be expected.

Theorem 9 *For all $n \geq 1$ and all allocation strategies ensuring that for all distributions $\nu_1, \ldots, \nu_K$ over the rewards, $N_{j,n} \geq \ln n$, for all arms j, the simple regret is bounded by*

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

Proof: We start as in the proof of Theorem 7 by writing

$$\mathbb{E}r_n \leq \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \widehat{\mu}_{j^*,n}\big\}$$

and upper bound the probabilities for all j by using the union bound,

$$\begin{aligned}
&\mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \widehat{\mu}_{j^*,n}\big\} \\
&\quad \leq \quad \mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta_j/2 \text{ or } \widehat{\mu}_{j^*,n} \leq \mu_{j^*} - \Delta_j/2\big\} \\
&\quad \leq \quad \mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta_j/2\big\} + \mathbb{P}\big\{\widehat{\mu}_{j^*,n} \leq \mu_{j^*} - \Delta_j/2\big\} \\
&\quad \leq \quad \frac{4}{\Delta_j^2}\left(\frac{1}{n}\right)^{\Delta_j^2/2} \, ,
\end{aligned} \tag{5}$$

where we now prove the last inequality; the arguments in the proof being symmetric, we only show that for all $j =$

$1, \ldots, K$ and all $\Delta > 0$,

$$\mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta\big\} \leq \frac{1}{2\Delta^2}\left(\frac{1}{n}\right)^{2\Delta^2} .$$

We use the assumption on the $N_{j,n}$ and resort, again, to the union bound,

$$\begin{aligned}
&\mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta\big\} \\
&\quad\leq \mathbb{P}\big\{\exists\, t \in [\![\, \lceil \ln n \rceil,\, n]\!] \text{ s.t. } N_{j,n} = t \\
&\qquad\qquad \text{and } X_{j,1} + \ldots + X_{j,t} - t\mu_j \geq t\Delta\big\} \\
&\quad\leq \sum_{t=\lceil \ln n \rceil}^{n} \mathbb{P}\big\{X_{j,1} + \ldots + X_{j,t} - t\mu_j \geq t\Delta\big\} \\
&\quad\leq \sum_{t=\lceil \ln n \rceil}^{n} \exp\left(-\frac{2t^2\Delta^2}{t}\right) = \sum_{t=\lceil \ln n \rceil}^{n} \exp\left(-2t\Delta^2\right) \\
&\quad\leq \int_{\ln n}^{\infty} \exp\left(-2\Delta^2 t\right)\, \mathrm{d}t \\
&\quad= \frac{1}{2\Delta^2}\exp\left(-2\Delta^2 \ln n\right) = \frac{1}{2\Delta^2}\left(\frac{1}{n}\right)^{2\Delta^2} ,
\end{aligned}$$

where we used Hoeffding's inequality (for i.i.d. random variables, see Hoeffding, 1963) for the third inequality, as in the proof of Theorem 7. ■

4.2.2 EXP3 as allocation strategy

It is easy to see that the following strategy, called EXP3 with a mixing step, has a cumulative regret $\mathbb{E}R_n$ not more than something of the order of $\sqrt{nK \ln L}$, just like EXP3 without the mixing step. φ_1 is the uniform distribution and for $t \geq 2$, we define φ_t component-wise as

$$\varphi_{j,t} = (1 - \gamma_t)\frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{j,s}\right)}{\sum_{k=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{k,s}\right)} + \frac{\gamma_t}{K}$$

for all $j = 1, \ldots, K$, where, as in Section 3.1,

$$\eta_t = \sqrt{\frac{2\ln K}{K\, t}} \quad \text{and} \quad \widehat{\ell}_{k,s} = \frac{1 - X_{k,N_{k,s}}}{\varphi_{k,s}}\mathbb{I}_{\{I_s = k\}}$$

are the estimated losses associated to the rewards and $\gamma_t = \gamma\sqrt{(K \ln K)/t}$ for some $\gamma > 0$ is the (time-varying) mixing parameter.

Of course the cumulative regret depends linearly upon γ and the theoretical optimal choice is $\gamma = 0$ (as in Theorem 3). But for $\gamma > 0$, the mixing step ensures that (by concentration-of-the-measure arguments) each arm is sampled at least $\gamma_1 + \ldots + \gamma_n = \Theta\big(\sqrt{nK \ln K}\big)$ times, and thus the estimation of the means is fine at a mild (exponential in $\sqrt{n}$) error factor. The same argument as in the proof of Theorem 7 then concludes at a simple regret decreasing at an exponential in $\sqrt{n}$ rate whenever $\gamma > 0$, a rate to be compared to the polynomial rate proposed in Theorem 3 for the case $\gamma = 0$ (and the policies given by moving averages or empirical distributions).

The precise statement of the bound and its proof are omitted from this extended abstract due to lack of space (and because of the poor practical performances of EXP3 strategies with mixing step, see the comments for Figure 4).

5 Distribution-free and minimax bounds

5.1 Distribution-free upper bounds

For EXP3 sampling

In Section 4, we exhibited distribution-dependent bounds (i.e., bounds that may depend on the underlying distributions ν_j, usually through the gaps Δ_j). We now turn to distribution-free bounds on the simple regrets. They are of the form

$$\sup_{\nu_1, \ldots, \nu_K} \mathbb{E}r_n \leq B_{K,n}$$

where the supremum is taken over all K-tuples of probability distributions over $[0, 1]$. Theorem 3 indicates for instance that the variant of EXP3 considered there is such that $B_{K,n} = 4\sqrt{(K \ln K)/n}$.

Because of the form of the distribution-dependent bounds on the regret, it is easy to derive distribution-free bounds from them, by optimizing the bound in the gaps Δ_j as illustrated below. This is in contrast with the (statistical) problem of identifying the best arm, for which no non-trivial distribution-free bounds can be exhibited, since, intuitively, this problem gets arbitrarily complicated as some of the gaps Δ_j tend to 0. Here however, because the regret equals the product of the gaps and the probabilities of incorrect selection, facing small gaps helps. This illustrates once again the usual differences between statistical and learning problems.

For uniform sampling

The following corollary is a simple consequence of Theorem 7, via a worst-case study of the bound as a function of the Δ_j (see the straightforward proof in appendix).

Corollary 10 *The uniform sampling allocation associated to the empirical successes policy considered in Section 4.1 ensures that the simple regrets are uniformly bounded by*

$$\sup_{\nu_1, \ldots, \nu_K} \mathbb{E}r_n \leq e^{-1/2}\frac{K - 1}{\sqrt{\lfloor n/K \rfloor}} = \Theta\left(\frac{K\sqrt{K}}{\sqrt{n}}\right)$$

for all $n \geq K$.

For UCB1 sampling

We now want to perform the same optimization to get distribution-free bounds for either of the sequences of empirical distributions and moving averages policies (1) based on the allocation strategy of UCB1. The bound for UCB1 we recalled in Theorem 2 cannot be optimized directly, for it tends to infinity as one of the Δ_j tends to 0. However, a simple modification of the proof of Auer et al. (2002a, Theorem 1) leads, for the very same forecaster UCB1, to a suitable bound on $\mathbb{E}R_n$ that in turn, still thanks to Lemma 1, implies a bound on $\mathbb{E}r_n$.

Theorem 11 *For all probability distributions $\nu_1, \ldots, \nu_K$ on $[0, 1]$,* UCB1 *ensures that its expected regret is bounded as*

$$\mathbb{E}R_n \leq (K - 1)\sqrt{n\left(8\ln n + 1 + \frac{\pi^2}{3}\right)}$$

for all $n \geq 1$.

Thus the expected simple regrets of either of the sequences of empirical distributions and moving averages (1) based on the allocation strategy of UCB1 *are uniformly bounded as*

$$\mathbb{E}r_n \leq (K-1)\sqrt{\frac{8\ln n + 1 + \frac{\pi^2}{3}}{n}}$$

for all $n \geq 1$.

Proof: It can be extracted from the proof of Auer et al. (2002a, Theorem 1) that for all suboptimal arm j,

$$\mathbb{E}N_{j,n} \leq \frac{8\ln n}{\Delta_j^2} + 1 + \frac{\pi^2}{3}\ ;$$

on the other hand, the simple upper bound $\mathbb{E}N_{j,n} \leq n$ always holds true. Therefore, $\mathbb{E}N_{j,n}$ is less than the minimum of these two bounds, and hence, less than the geometric mean of them,

$$\mathbb{E}N_{j,n} \leq \sqrt{n\left(\frac{8\ln n}{\Delta_j^2} + 1 + \frac{\pi^2}{3}\right)}\ .$$

The first bound of the theorem now follows from the very definition of the regret,

$$\begin{aligned}\mathbb{E}R_n &= \sum_{j:\Delta_j>0} \Delta_j\, \mathbb{E}N_{j,n} \\ &\leq \sum_{j:\Delta_j>0} \sqrt{n\left(8\ln n + \Delta_j^2\left(1+\frac{\pi^2}{3}\right)\right)}\ . \quad (6)\end{aligned}$$

■

With a somewhat refined analysis, one can indeed compute the exact order of magnitudes of $\mathbb{E}R_n$ in K and n for UCB1,

$$\mathbb{E}R_n = \Theta\left(K\sqrt{n\ln n}\right)\ ,$$

which, still by virtue of Lemma 1, shows that either of the sequences of empirical distributions and moving averages (1) based on the allocation strategy of UCB1 has simple regrets of the order of

$$\mathbb{E}r_n = \Theta\left(K\sqrt{\frac{\ln n}{n}}\right)\ .$$

The details are omitted from this extended abstract.

We now turn to the combination of UCB1 as allocation strategy and empirical successes as final policy, as studied in Section 4.2.1. The bound of Theorem 9 cannot be optimized over the Δ_j to yield a non-trivial distribution-free bound. One way around is to use a sharper concentration inequality, namely Bernstein's inequality for martingales, see, e.g., Freedman (1975) or Cesa-Bianchi et al. (2005, Lemma 15), and partition the set of possible values for the $N_{j,n}$ in a finer way (in not more than something of the order of $\ln n$ bins).

We could also have followed the approach of the proof of Theorem 11 and combined (5) with the fact that the probabilities of interest are always less than 1 to get, via Hölder's averages, the bound

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

for all $\alpha \in [0,1]$. Choosing $\alpha > 0$ arbitrarily close to 0 would get the distribution-free convergence rate for the simple regrets arbitrarily close to $1/\sqrt{\ln n}$, but with a leading constant by far worse than in the bound stated below. Recall that the condition on the number of times each arm is pulled is natural in view of the result of Lemma 8 and the comments following it.

Theorem 12 *For all* $n \geq 1$ *and all allocation strategies ensuring that for all distributions* $\nu_1, \ldots, \nu_K$ *over the rewards,* $N_{j,n} \geq \log_2 n$ *for all arms* j*, the simple regret is bounded by*

$$\mathbb{E}r_n \leq \frac{4\sqrt{2}\,(K-1)}{\sqrt{2}-1}\ \frac{1}{\sqrt{\log_2 n}}\ .$$

Proof: The proof is a variation on the one of Theorem 9; we essentially replace Hoeffding's inequality (for i.i.d. random variables) by the sharper Bernstein's inequality (for martingales). We simply prove below that for all $j = 1, \ldots, K$ and all $\Delta > 0$,

$$\mathbb{P}\big\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta\big\} \leq \frac{\sqrt{2}}{\Delta\big(\sqrt{2}-1\big)\sqrt{\log_2 n}}\ ; \quad (7)$$

and the conclusion will follow (since here also the argument is symmetric for j and j^*).

The martingale difference sequence we consider here is formed by the $(Y_t - \mu_j)\mathbb{I}_{\{I_t=j\}}$ (bounded by 1), with $t = 1, \ldots, n$, where we recall from Figure 1 that Y_t is the reward obtained by the forecaster at round t. The associated martingale $M_{j,n}$ and sum of conditional variances $V_{j,n}$ are given by

$$M_{j,n} = N_{j,n}\,(\widehat{\mu}_{j,n} - \mu_j) \quad \text{and} \quad V_{j,n} = \mu_j(1-\mu_j)\,N_{j,n}\ .$$

Using that $x(1-x) \leq 1/4$ for all $[0,1]$, we have, for all $\varepsilon > 0$ and $v > 0$,

$$\begin{aligned}&\mathbb{P}\left\{\widehat{\mu}_{j,n} \geq \mu_j + \frac{\varepsilon}{N_{j,n}} \text{ and } N_{j,n} \leq 4v\right\} \\ \leq\ &\mathbb{P}\left\{\widehat{\mu}_{j,n} \geq \mu_j + \frac{\varepsilon}{N_{j,n}} \text{ and } N_{j,n} \leq \frac{v}{\mu_j(1-\mu_j)}\right\} \\ =\ &\mathbb{P}\big\{M_{j,n} \geq \varepsilon \text{ and } V_{j,n} \leq v\big\} \\ \leq\ &\exp\left(-\frac{\varepsilon^2}{2(v+\varepsilon/3)}\right)\ , \quad (8)\end{aligned}$$

where the last step is exactly the statement of Berstein's inequality. We partition the set of possible values for $N_{j,n}$ into the (integer) intervals $[\![2^r,\ 2^{r+1}-1]\!]$, for r varying from

$r_0 = \lfloor \log_2 \lceil \log_2 n \rceil \rfloor$ to $r_1 = \lfloor \log_2 n \rfloor$,

$$
\begin{aligned}
&\mathbb{P}\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta\} \\
&\quad = \sum_{r=r_0}^{r_1} \mathbb{P}\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta \text{ and } 2^r \leq N_{j,n} \leq 2^{r+1} - 1\} \\
&\quad \leq \sum_{r=r_0}^{r_1} \mathbb{P}\left\{\widehat{\mu}_{j,n} \geq \mu_j + \frac{2^r \Delta}{N_{j,n}} \text{ and } N_{j,n} \leq 2^{r+1}\right\} \\
&\quad \leq \sum_{r=r_0}^{r_1} \exp\left(-\frac{(2^r\Delta)^2}{2\,(2^{r-1} + (2^r\Delta)/3)}\right) ,
\end{aligned}
$$

where the last step is an application of (8) with $v = 2^{r-1}$ and $\varepsilon = 2^r\Delta$. Substituting $2^{r-1} + (2^r\Delta)/3 \leq 2^r$ yields

$$
\begin{aligned}
&\mathbb{P}\{\widehat{\mu}_{j,n} \geq \mu_j + \Delta\} \\
&\quad \leq \sum_{r=r_0}^{r_1} \exp\left(-\frac{(2^r\Delta)^2}{2\,(2^{r-1} + (2^r\Delta)/3)}\right) \leq \sum_{r=r_0}^{r_1} e^{-2^r\Delta^2/2} \\
&\quad \leq M \sum_{r=r_0}^{r_1} \frac{1}{\Delta\sqrt{2^r}} \leq \frac{e^{-1/2}}{\Delta\left(\sqrt{2}\right)^{r_0}} \frac{1}{1 - 1/\sqrt{2}} ,
\end{aligned}
$$

where we used for the third inequality that $M = e^{-1/2}$ is the maximum of $x \in [0, \infty[\mapsto x\, e^{-x^2/2}$ and in the last step, that we are left with a geometric sum. The proof of (7) is concluded by noting that

$$
\begin{aligned}
\left(\sqrt{2}\right)^{r_0} &= \exp\left(\frac{1}{2}\lfloor \log_2 \lceil \log_2 n \rceil \rfloor\right) \\
&\geq e^{-1/2} \exp\left(\frac{1}{2}\log_2 \log_2 n\right) = e^{-1/2}\sqrt{\log_2 n} .
\end{aligned}
$$

■

5.2 Minimax bounds

We presented in Section 3.2 some distribution-dependent lower bounds on the simple regrets and we now focus on distribution-free such bounds. We show below that the orders of magnitude in the number of rounds n are different for two types of bounds. Theorems 6 and 7 indicate that the optimal order of magnitude for distribution-dependent bounds is exponential, whereas it is $1/\sqrt{n}$ for distribution-free bounds as asserted by Theorems 3 and 13. A similar situation arises for the cumulative regret, see Lai and Robbins, 1985 (optimal $\ln n$ rate for distribution-dependent bounds) versus Auer et al., 2002b (optimal $\sqrt{n}$ rate for distribution-free bounds).

Theorem 13 *For all $n \geq 1$ and $K \geq 2$ such that $n > K/(4\ln(4/3))$, the simple regrets are bounded in a minimax sense as*

$$
\inf \sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \geq \frac{\sqrt{2} - 1}{2\sqrt{2\ln(4/3)}} \sqrt{\frac{K}{n}}
$$

where the infimum is taken over all (randomized) allocation strategies and all associated policies.

The proof is almost the same as the proofs of the lower bounds on the cumulative regret in multi-armed bandit problems, see Auer et al. (2002b, Appendix A) and Cesa-Bianchi and Lugosi (2006, Section 6.9), except that we only have to consider the final round here. It can be found in the appendix.

Actually, in view of Lemma 1, the way the proofs could go for cumulative regrets is by lower bounding the simple regrets and use that the minimax value for cumulative regrets after n rounds is at least n times the minimax value of simple regrets.

It is an open question whether the extra $\sqrt{\ln K}$ factor of Theorem 3 is necessary or if the lower bound of Theorem 13 should be improved. $\mathbb{E}r_n$ is usually easier to handle than $\mathbb{E}R_n$. Fancier techniques than those of the proof of Theorem 13 might lead to an improvement of the lower bound on the minimax value of $\mathbb{E}r_n$ (and thus, on the one of $\mathbb{E}R_n$); one might take inspiration, for instance, from the exact computation of the minimax value of simple regret $\mathbb{E}r_n$ for two-armed bandits in Schlag (2006), where the minimax optimal strategy is exhibited. The latter, called binomial average rule, samples the two arms equally often and basically chooses the one with best empirical average at even rounds (and a slight adaptation of that at odd rounds, when one arm has been sampled once less than the other one). He however points out that the minimax strategy is probably more complicated whenever there are at least three arms. (A related result of this paper is that in Corollary 10 we have a worse dependence in K for the uniform sampling together with empirical successes than in Theorem 3.)

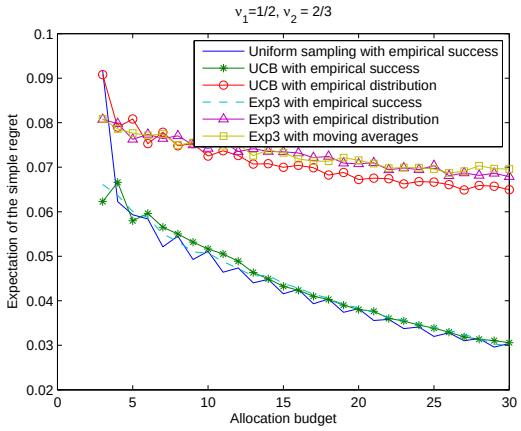

Figure 2: $K = 2$ arms with Bernoulli distributions with parameters $1/2$ and $2/3$.

6 Simulation study

Figures 2–4 present some experimental results on artificial data. We considered three different allocation strategies (uniform sampling, EXP3 with and without mixing parameter γ, and UCB1) and three associated policies (empirical distributions, moving averages, and empirical successes). The corresponding simple regrets are computed over 10 000 runs of each K–tuple of distributions and we plot their averages, which approximate well the expectations $\mathbb{E}r_n$. The distributions used for the simulations are given by Bernoulli distributions, in number and with parameters depending on the experiment (see the captions of the different figures for a de-

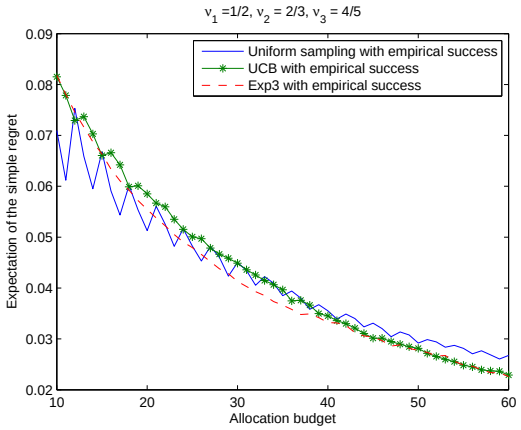

Figure 3: $K = 3$ arms with Bernoulli distributions with parameters $1/2$, $2/3$, and $4/5$.

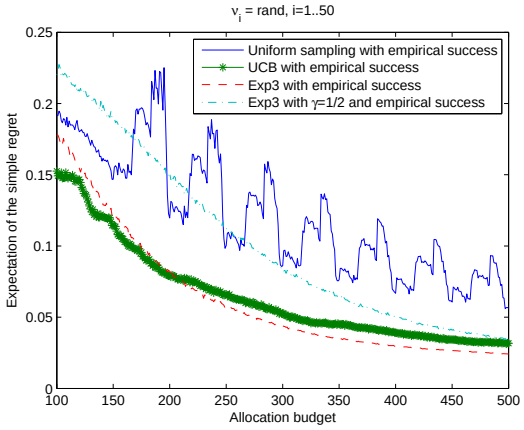

Figure 4: $K = 50$ arms with Bernoulli distributions; all parameters chosen independently at random in $[0, 1]$.

scription of each K–tuple). Though we only offer a limited number of graphical illustrations of the performances, we mention that all situations illustrated below are typical and do not result from a particular choice of the underlying distributions.

Figure 2 essentially shows that empirical successes policies described and studied in Section 4 clearly outperform the policies (empirical distributions and moving averages) of Section 3.1, even though the theoretical rates of convergence of the latter are usually better than those of the former. (We recall that since UCB1 is a deterministic strategy, empirical distributions and moving averages coincide for it, see the comments before Theorem 2. Here, we take a mixing parameter $\gamma = 0$ for EXP3, as in Section 3.1.) Empirical distributions and moving averages seem to be of theoretical interest only; they probably suffer in practice from being too conservative.

Figure 3 shows that for a small number of arms (typically, for K less than 10) the empirical successes based on uniform sampling, UCB1 sampling, and EXP3 sampling (with no mixing parameter) have comparable performances, whereas Figure 4 (see also Figure 5 in the appendix) illustrates that this is not the case anymore for larger numbers (typically, more than 10). In this case, UCB1 is good for small values of n and EXP3 is the best allocation strategy when n is larger. A closer look to the small values of K indicates that whereas uniform sampling is the best allocation strategy for $K = 2$, it is never the optimal allocation strategy for $K \geq 3$; this is not even due to the fact that we discard some final rounds to compute the empirical successes, as the reader may notice that at rounds with indexes multiple of K the performances of this strategy is always off the others. Actually, the last rounds represented in Figure 3 show that the simple regret associated to the uniform sampling is twice larger than those associated to UCB1 and EXP3 samplings. This ranking may be surprising at first sight, since, in theory, the distribution-dependent rates for uniform sampling combined to empirical successes indicate an exponential decrease of the simple regrets (Theorem 7), whereas Theorem 4 together with Theorem 2 shows that the simple regrets associated to UCB1 and EXP3, even through empirical successes, cannot converge faster than at a polynomial rate. In principle, the uniform allocation should thus take the lead over it, a phenomenon we only observed for numbers of rounds n so large that simple regrets are smaller than 10^{-10}, a precision for which little can be guaranteed in terms of correct numerical computations.

Figure 4 studies also whether a mixing parameter γ would benefit to EXP3–based strategies. We only reported one value but tested many; in all cases, the performances of the mixing–EXP3 strategies interpolated the ones of uniform sampling and EXP3 without mixing, therefore performing worse than the latter for more than $K \geq 3$ arms. In Section 4.2 and Figure 4, we considered time-varying mixing parameters, but it turns out that that on simulations not reported here, the use of constant mixing parameters does not change the picture: it is always better not to mix the EXP3 distributions with the uniform distribution.

The simulations are somewhat disappointing in the sense that the theoretical best strategies are not the best in practice; but they point out to which extend simple regrets can gain from being computed with non-uniform allocation strategies, a result well-known in the Gaussian case (see Chen et al., 2000) but not, to the best of our knowledge, in the case of Bernoulli distributions. That EXP3 or UCB1 be the best allocation strategies also show, surprisingly enough, the interest of exploration–exploitation trade-offs for pure exploration problems!

7 Conclusion and open problems

We showed in this paper that even for the pure exploration problem, the exploration–exploitation trade-off is useful, via forecasters like UCB1 and EXP3. Together with the moving averages policy, they lead to good, and even almost optimal, distribution-free bounds; associated to the empirical successes policy, they show interesting practical performances. These results are somewhat surprising in view of the distribution-dependent bounds that indicate that forecasters performing good exploration–exploitation trade-offs in terms of cumulative regrets have simple regrets with orders of mag-

nitude way off those of some more naive forecasters, as, for instance, the uniform sampling together with empirical successes policy.

Three extensions are left for future research. The first would be to take into account that getting the reward of an arm might take a (random) time that depends on the arm, to model, e.g., that some paths are more complicated to evaluate in the motivating example of tree search; this was done for cumulative regret by György et al. (2007).

The second is to study in detail the case of probability distributions over the rewards that are not compactly supported and check whether the asserted links between cumulative and simple regrets also hold. The prototypical case is of course the case of Gaussian distributions, as in Chen et al. (2000).

Pure exploration for bandit problems in topological spaces

The third extension if of theoretical interest. For an interval I of $\mathbb{R}$, we denote by $\mathcal{P}_B(I)$ the set of probability distributions over I with first moments less than B. Given a topological space $\mathcal{X}$, we call environment on $\mathcal{X}$ any mapping $M : \mathcal{X} \to \mathcal{P}_B(I)$ (for some B and I). We say that M is continuous if the mapping that associates to each $x \in \mathcal{X}$ the expectation $\mu(x)$ of $M(x)$ is continuous.

The $\mathcal{X}$–armed bandit problem is as follows. An environment M on $\mathcal{X}$ is fixed by Nature. The forecaster may choose at each round a point I_t in $\mathcal{X}$ and gets a rewards distributed according to $M(I_t)$. We say that a family $\mathcal{F}$ of environments is explorable–exploitable (respectively, explorable) if for any environment $M \in \mathcal{F}$, the forecaster can guarantee that his expected per-round reward converges to the expectation μ^* of the best distribution among the $M(x)$ (respectively, recommends a random point of $\mathcal{X}$ such that its associated expected reward in a one-shot new instance of the problem is close to μ^*). Explorability is a milder requirement than explorability–exploitability, as can be seen by an equivalent of Lemma 1 in this setting.

This paper was about the family of all environments over $\mathcal{X} = \{1, \dots, K\}$ and a fixed bounded I being explorable. By using the doubling trick, it can be seen that for $I = \mathbb{R}$, a fixed bound B, and a separable space $\mathcal{X}$, the family of all continuous environments defined with these parameters is explorable. On the negative side, one can show that if $\mathcal{X}$ is uncountable, there exists a family of environments over $I = [0, 1]$ (with $B = 1$) that is not explorable. In addition to investigating this further, one natural question is to see whether there are situations where explorability is possible but not explorability–exploitability.

References

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, pages 229–243, 2006.

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.

D.A. Berry and B. Fristedt. *Bandit Problems*. Chapman and Hall, 1985.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, UK, 2006.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE: Transactions on Information Theory*, 51:2152–2162, 2005.

C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3): 251–270, 2000.

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

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

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.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.

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

A. György, L. Kocsis, I. Szabó, and C. Szepesvári. Continuous time associative bandit problems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 830–835, 2007.

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

A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Gap-free bounds for stochastic multi-armed bandit. In *Proceedings of the 17th International Federation of Automatic Control World Congress*, 2008.

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.

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

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

T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

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

G. Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris-Sud, Orsay, France, May 2005.

A Omitted proofs and omitted figure

We include here all material omitted in the main text for the sake of the limitation on the number of pages. They will be dropped in the final submission if the paper is accepted. Reviewers may of course advise us to put back some of the following proofs and recommend to omit some other sections of the paper instead.

A.1 Proof of the second statement of Theorem 7

Proof: We start by writing

$$\begin{aligned}\mathbb{E}r_n &= \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}\Phi_{j,n} \\ &\leq \left(\max_{j=1,\ldots,K} \Delta_j\right) \mathbb{P}\left\{\max_{j:\Delta_j>0} \widehat{\mu}_{j,n} \geq \widehat{\mu}_{j^*,n}\right\}\end{aligned}$$

where the second 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

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

for some function f, where for all $t = 1, \ldots, \lfloor n/K \rfloor$, we denote by $\vec{X}_t$ the vector $(X_{1,t}, \ldots, X_{K,t})$. ($f$ is defined as a maximum of at most $K-1$ sums of differences.) It is straightforward that since all random variables of interest take values either 0 or 1 here, the bounded differences condition is satisfied with ranges all equal to 2. Therefore, the indicated concentration inequality states that

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

for all $\varepsilon > 0$. We choose

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

(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

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

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 $j = 1, \ldots, K$. Independence and Hoeffding's lemma (see, e.g., Devroye and Lugosi, 2001, Chapter 2) imply that for all $\lambda > 0$,

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

(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, Devroye and Lugosi, 2001, Chapter 2) then yields

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

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

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

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

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

The result follows for n such that

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

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

A.2 Proof of Lemma 8

The following proof is close to the one of Audibert et al. (2008, Proposition 1).

References

J-Y. Audibert, R. Munos, and Cs. Szepesvári. Tuning bandit algorithms in stochastic environments. *Theoretical Computer Science*, 2008. To appear.

Proof: The second part of the lemma follows by straightforward calculations. To prove the first part, for integers $t \geq 2$, we denote, by $p(t)$ the unique solution of the equation

$$\sqrt{\frac{2\ln t}{x}} = 1 + \sqrt{\frac{2\ln t}{t-x}} \, .$$

Simple calculations (not reported in this extended abstract) show that for $t \geq 2$,

$$p(t+1) \leq p(t) + \frac{1}{2} \tag{9}$$

$$\text{and} \qquad 2\ln\left(t\left(1 - \sqrt{\frac{2\ln t}{t}}\right)\right) \leq p(t) \leq \frac{t}{2} \, . \tag{10}$$

We fix an arm, say $j = 1$, and prove by induction that $N_{1,t} \geq p(t-1)$ for all $t \geq 3$; the lemma then follows by the first inequality in (10). Since UCB1 pulls, by definition, each

arm once before using the rule (2), we have that $N_{1,3} \geq 1 \geq p(2)$, where we used the upper bound on $p(t)$ given in (10). Assume now that for some $t \geq 3$, we have $N_{1,t} \geq p(t-1)$. If $I_{t+1} = 1$, then $N_{1,t+1} = N_{1,t} + 1 \geq p(t-1) + 1 \geq p(t)$, where we used (9). If on the contrary $I_{t+1} = 2$, then this is because

$$\widehat{\mu}_{1,t} + \sqrt{\frac{2\ln t}{N_{1,t}}} \leq \widehat{\mu}_{2,t} + \sqrt{\frac{2\ln t}{n - N_{1,t}}} \; ;$$

in particular,

$$\sqrt{\frac{2\ln t}{N_{1,t}}} \leq 1 + \sqrt{\frac{2\ln t}{n - N_{1,t}}} \; ,$$

revealing that $N_{1,t} \geq p(t)$. In this case, we thus also have $N_{1,t+1} = N_{1,t} \geq p(t)$. ■

A.3 Proof of Corollary10

Proof: In view of Theorem 7, since the gaps Δ_j all lie in $[0,1]$ and at least one of them equals 0, the maximum of the function

$$(x_1, \ldots, x_{K-1}) \in [0,1]^{K-1} \mapsto \sum_{j=1}^{K-1} x_j e^{-x_j^2 \lfloor n/K \rfloor / 2}$$

is a suitable minimax bound. By separation of the variables (and since it only helps to take $x_j > 0$), this maximum is $K-1$ times the maximum of

$$g : x \in]0,1] \mapsto x e^{-x^2 \lfloor n/K \rfloor / 2} \; .$$

The latter is identified as $g\big(\sqrt{1/\lfloor n/K \rfloor}\big)$ by considering $\ln g$, which has first derivative $1/x - x\lfloor n/K \rfloor$, vanishing at $x = \sqrt{1/\lfloor n/K \rfloor}$, and negative second derivative. ■

A.4 Proof of Theorem 13

Proof: We introduce first $\mathbb{P}_0$ and $\mathbb{E}_0$ as the probability distribution and expectation associated to the K–tuple of symmetric Bernoulli distributions $\nu_j = \mathcal{B}(1/2)$ for all $j = 1, \ldots, K$. We fix $0 \leq \varepsilon \leq 1/4$. For all $i = 1, \ldots, K$, we also denote by $\mathbb{P}_i$ and $\mathbb{E}_i$ those associated to the K–tuple given by $\nu_i = \mathcal{B}(1/2 + \varepsilon)$ and $\nu_j = \mathcal{B}(1/2)$ for all $j \neq i$.

We fix first a deterministic forecaster. The distributions φ_t and Φ_n are given by Dirac masses on points that we denote by I_t, the indexes of the pulled arms, and by J_n, the recommended action. We have

$$\sup_{\nu_1, \ldots, \nu_K} \mathbb{E} r_n \geq \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_i r_n = \frac{\varepsilon}{K} \sum_{i=1}^{K} \left(1 - \mathbb{P}_i \left\{J_n = i\right\}\right)$$

where the last equality comes from the fact that we suffer an expected regret of ε whenever we did not recommend the optimal action (with index i under $\mathbb{P}_i$). Denoting by $\mathbb{P}^J$ the image distribution of some probability distribution $\mathbb{P}$ by a random variable J, we then have, by Pinsker's inequality, for all $i = 1, \ldots, K$,

$$\begin{aligned} \mathbb{P}_i \left\{J_n = i\right\} - \mathbb{P}_0 \left\{J_n = i\right\} \; &\leq \; \sqrt{\frac{1}{2} \mathcal{K}\left(\mathbb{P}_0^{J_n}, \mathbb{P}_i^{J_n}\right)} \\ &\leq \; 2\sqrt{\ln(4/3)}\, \varepsilon \sqrt{\sum_{t=1}^{n} \mathbb{E}_0 N_{i,n}} \end{aligned}$$

where the last inequality proceeds from an inequality on Kullback-Leibler divergences stated in Cesa-Bianchi and Lugosi (2006, top of page 168). Averaging over i and using concavity of the root together with $N_{1,n} + \ldots + N_{K,n} = n$, we get

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{P}_i \left\{J_n = i\right\} \leq \frac{1}{K} + 2\sqrt{\ln(4/3)}\, \varepsilon \sqrt{\frac{n}{K}} \; .$$

Substituting this inequality, we have proved

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_i r_n \geq \varepsilon \left(1 - \frac{1}{K} - 2\sqrt{\ln(4/3)}\, \varepsilon \sqrt{\frac{n}{K}}\right) \; .$$

The proof is concluded as in Cesa-Bianchi and Lugosi (2006, Section 6.9), first by optimizing over ε and then by considering the case of randomized policies, which follows from the bound for deterministic strategies basically by taking expectations with respect to the auxiliary randomizations the forecaster has access to, see the reference above for more details. ■

A.5 Additional figure for the simulation study in Section 6

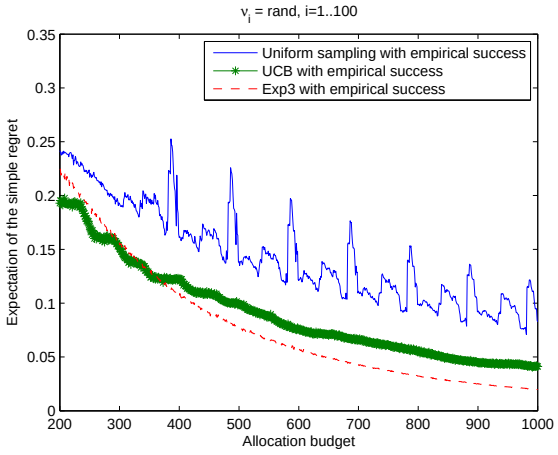

Figure 5: $K = 100$ arms with Bernoulli distributions; all parameters chosen independently at random in $[0,1]$.