Information complexity in bandit subset selection

Emilie Kaufmann[1], Shivaram Kalyanakrishnan[2]

[1] Institut Mines-Telecom; Telecom ParisTech kaufmann@telecom-paristech.fr

[2] Yahoo Labs! Bangalore shivaram@yahoo-inc.com

1 Introduction

Résumé : We consider the problem of efficiently exploring the arms of a stochastic bandit to identify the best subset. For the Explore-m problem, we derive improved bounds by using KL-divergence-based confidence intervals. Whereas the application of a similar idea in the regret setting has yielded bounds in terms of the KL-divergence between the arms, our bounds in the pure-exploration setting involve the "Chernoff information" between the arms. In addition to introducing this novel quantity to the bandits literature, we contribute a comparison between strategies based on "uniform sampling" and "adaptive sampling" for pure-exploration problems, finding evidence in favor of the latter.

Mots-clés : bandit model, pure-exploration, subset selection

We consider a stochastic bandit model with a finite number of arms $K \geq 2$. Each arm a corresponds to a Bernoulli distribution with mean p_a ; the arms are numbered such that $p_1 \geq p_2 \geq ... \geq p_K$. Each draw of arm a yields a reward drawn from an unknown distribution $\mathcal{B}(p_a)$. In the classical "regret" setting, an agent seeks to sample arms sequentially in order to maximize its cumulative reward, or equivalently, to minimize its regret. This setting was originally motivated by clinical trials (Thompson, 1933) wherein the number of subjects cured is to be maximized through the judicious allocation of competing treatments. By contrast, the "pure exploration" setting models an *off-line* regime in which the rewards accrued while learning are immaterial ; rather, the agent has to identify an optimal set of m arms at the end of its learning (or *exploration*) phase. Such a setting would naturally suits a company that conducts a dedicated testing phase for its products to determine which m products to launch into the market. Bubeck *et al.* (2011) present an informative comparison between the regret and pure-exploration settings.

In this paper, we consider the pure-exploration problem of finding the m best arms, introduced by Kalyanakrishnan & Stone (2010) as "Explore-m". This problem, which generalizes the single-arm-selection problem studied by Even-Dar *et al.* (2006), is as follows. For some fixed tolerance $\epsilon \in [0, 1]$, let $\mathcal{S}^*_{m,\epsilon}$ be the set of all (ϵ, m)-optimal arms : the set of arms a such that $p_a \geq p_m - \epsilon$. Observe that the set of m best arms, $\mathcal{S}^*_m = \{1, 2, \ldots, m\}$, is necessarily a subset of $\mathcal{S}^*_{m,\epsilon}$. For a given mistake probability $\delta \in]0, 1]$, our goal is to design an algorithm that after using a finite (but possibly random) total number of samples from the arms $\mathcal{N}$ returns $\mathcal{S}_\delta$, a set of m arms satisfying $\mathbb{P}(\mathcal{S}_\delta \subset \mathcal{S}^*_{m,\epsilon}) \geq 1 - \delta$. We desire $\mathcal{N}$ to be small in expectation.

In the regret setting, a recent line of research has yielded essentially optimal algorithms. While the regret bound for the UCB algorithm of Auer *et al.* (2002) is optimal in its logarithmic dependence on the horizon, its accompanying problem-specific constant does not match the lower bound provided by Lai & Robbins (1985). If $N_a(n)$ denotes the number of draws of arm a up to time n, Lai & Robbins' result states that every consistent algorithm (i.e. whose regret R_n is such that $\forall \alpha > 1,\ R_n = o(n^\alpha)$ on every bandit problem),

$$p_a < p_1 \ \Rightarrow \ \liminf_{n\to\infty} \frac{\mathbb{E}[N_a(n)]}{\log(n)} \geq \frac{1}{d(p_a, p_1)},$$

with $N_a(n)$ the number of draws of arm a up to time n and $d(x, y)$ the Kullback-Leibler divergence between two Bernoulli distributions :

$$d(x, y) = KL\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x \log\left(\frac{x}{y}\right) + (1 - x) \log\left(\frac{1 - x}{1 - y}\right).$$

Cappé *et al.* (2013) and references therein show that by replacing UCB's Hoeffding's inequality-based bounds with upper bounds based on Kullback-Leibler divergence, the constant, too, becomes optimal.

The primary contribution of this paper is a set of similarly-improved bounds for the pure-exploration setting. We show improvements by replacing Hoeffding-based bounds with KL-divergence-based bounds in corresponding algorithms. Interestingly, our analysis exposes key differences between the pure-exploration and regret settings : the improved sample-complexity bounds we obtain here involve the *Chernoff information* between the arms, and not KL-divergence as in the regret setting.

Algorithms for pure-exploration broadly fall into two categories : algorithms based on *uniform sampling and eliminations* (Even-Dar *et al.* (2006); Heidrich-Meisner & Igel (2009)), and fully-sequential algorithms based on *adaptive sampling* (Kalyanakrishnan *et al.* (2012); Gabillon *et al.* (2012)). The second contribution of this paper is a comparison between these contrasting approaches, that share the use of lower and upper confidence bounds. We consider both "Hoeffding" and "KL" versions of these algorithms : in each case our theoretical and experimental results point to the superiority of the adaptive sampling heuristic.

This paper is organized as follows. We present generic versions of the Racing and LUCB algorithms in Section 2, proceeding to describe two specific instances, KL-Racing and KL-LUCB, that are analyzed in Section 3. We then discuss the complexity of Explore-m in Section 4. Section 5 presents corroborative results from numerical experiments, even in an alternative setting called Explore-m with fixed budget.

2 Two classes of algorithms based on Confidence Intervals

Just as upper confidence bounds have been used successfully in the regret setting, most existing algorithms for Explore-m have used both upper and lower confidence bounds on the means of the arms. We state here a generic version of an algorithm based on uniform sampling and eliminations, Racing, and a generic version of an algorithm based on adaptive sampling, LUCB. To describe these contrasting heuristics, we use generic confidence intervals, denoted by $\mathcal{I}_a(t) = [L_a(t), U_a(t)]$, where t is the round of the algorithm, $L_a(t)$ and $U_a(t)$ are the the lower and upper confidence bounds on the mean of arm a. Let $N_a(t)$ denote the number of draws, and $S_a(t)$ the sum of the rewards, gathered from arm a up to time t. Let $\hat{p}_a(t) = \frac{S_a(t)}{N_a(t)}$ be the corresponding empirical mean reward while $\hat{p}_{a,u}$ denotes the empirical mean of u i.i.d. rewards from arm a. Additionally, let $J(t)$ be the m arms with the highest empirical means at time t (for Racing, $J(t)$ only includes $m' \leq m$ arms if $m - m'$ have already been selected). Also, l_t and u_t are two 'critical' arms from $J(t)$ and $J(t)^c$ that are likely to be misclassified :

$$u_t = \underset{j \notin J(t)}{\operatorname{argmax}}\, U_j(t) \quad \text{and} \quad l_t = \underset{j \in J(t)}{\operatorname{argmin}}\, L_j(t).$$

2.1 An algorithm based on uniform sampling and eliminations : Racing

The idea of Racing dates back to Maron & Moore (1997), who introduced it in the context of model selection. It became the successive elimination algorithm of Even-Dar *et al.* (2006) for Explore-1. (using only rejects). The idea of using both accepts and rejects was then introduced by Heidrich-Meisner & Igel (2009) to a setting like Explore-m, applied within the context of reinforcement learning ; the authors do not formally analyze the algorithm's sample complexity, as we do here. More precisely, the Racing algorithm maintains a set of remaining arms $\mathcal{R}$, of selected arms $\mathcal{S}$ and of discarded arms $\mathcal{D}$. At round t this algorithm samples all the arms in $\mathcal{R}$ (i.e. samples uniformly the remaining arms), updates confidence intervals, computes the set $J(t)$ of empirical $m - |\mathcal{S}|$ best, the set $J(t)^c = \mathcal{R} \backslash J(t)$, critical arms u_t and l_t, and :

– selects the empirical best of $\mathcal{R}$, a_B if its LCB is bigger than all UCB's of all the arms in $J(t)^c$:

$$L_{a_B}(t) > U_{u_t}(t) - \epsilon \;\Rightarrow\; \mathcal{S} = \mathcal{S} \cup \{a_B\},\; \mathcal{R} = \mathcal{R} \backslash \{a_B\}$$

– discards the empirical worst of $\mathcal{R}$, a_W if its UCB is smaller than all the LCB's of all the arms in $J(t)$:

$$U_{a_W}(t) < L_{l_t}(t) + \epsilon \;\Rightarrow\; \mathcal{D} = \mathcal{D} \cup \{a_W\},\; \mathcal{R} = \mathcal{R} \backslash \{a_W\}$$

2.2 An algorithm based on adaptive sampling : LUCB

A general version of the LUCB algorithm proposed by Kalyanakrishnan *et al.* (2012) can be stated using generic confidence bounds U and L, while the original LUCB uses Hoeffding confidence regions. Unlike Racing, this algorithm does not sample the arms uniformly ; rather, it draws at each round the two critical

arms u_t and l_t. This adaptive *sampling strategy* is associated with the natural *stopping criterion* $(B(t) < \epsilon)$ where $B(t) := U_{u_t}(t) - L_{l_t}(t)$. The UGapE algorithm of Gabillon *et al.* (2012) is also an adaptive sampling algorithm, that uses an alternative definition of $J(t)$ using confidence bounds on the simple regret, and a correspondingly different stopping criterion $B(t)$.

The two algorithms mentioned above both use generic upper and lower confidence bounds on the mean of each arm, and one has the intuition that the smaller these confidence regions are, the smaller the sample complexity of these algorithms will be. Most of the previous algorithms use Hoeffding bounds. Mnih *et al.* (2008); Heidrich-Meisner & Igel (2009); Gabillon *et al.* (2012) has also considered the use of empirical Bernstein bounds, that can be tighter. In this paper, we introduce the use of confidence regions based on KL-divergence for Explore-m, inspired by recent improvements in the regret setting Cappé *et al.* (2013). We define, for some *exploration rate* $\beta(t, \delta)$,

$$u_a(t) := \max\{q \in [\hat{p}_a(t), 1] : N_a(t) d(\hat{p}_a(t), q) \leq \beta(t, \delta)\}, \text{ and} \quad (1)$$

$$l_a(t) := \min\{q \in [0, \hat{p}_a(t)] : N_a(t) d(\hat{p}_a(t), q) \leq \beta(t, \delta)\}. \quad (2)$$

Pinsker's inequality ($d(x, y) \geq 2(x - y)^2$) shows that KL-confidence regions are always smaller than those obtained with Hoeffding bounds, while one can show they share the same coverage probability :

$$\hat{p}_a(t) - \sqrt{\frac{\beta(t, \delta)}{2N_a(t)}} \leq l_a(t) \text{ and } u_a(t) \leq \hat{p}_a(t) + \sqrt{\frac{\beta(t, \delta)}{2N_a(t)}}. \quad (3)$$

We define, for a given function β, the **KL-Racing** and **KL-LUCB** algorithms with exploration rate β as the instances of Racing and LUCB, respectively, that use $u_a(t)$ and $l_a(t)$ as confidence bounds. In our theoretical and experimental analysis to follow, we address the "KL versus Hoeffding" and "adaptive sampling versus uniform sampling" questions.

3 Analysis of KL-Racing and KL-LUCB

Lemma 1 gives choices of β such that KL-Racing and KL-LUCB are correct with probability at least $1-\delta$. Note that we have he same guarantees for their Hoeffding counterpart, (Hoeffding)-Racing and LUCB.

Lemma 1
The (KL)-Racing algorithm using $\beta(t, \delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$, *with* $\alpha > 1$ *and* $k_1 > 1 + \frac{1}{\alpha - 1}$ *; and the (KL)-LUCB algorithm using* $\beta(t, \delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right) + \log\log\left(\frac{k_1 K t^\alpha}{\delta}\right)$, *with* $\alpha > 1$ *and* $k_1 > 2e + 1 + \frac{e}{\alpha - 1} + \frac{e+1}{(\alpha-1)^2}$, *are correct with probability at least* $1 - \delta$.

3.1 Chernoff information in Explore-m

Surprisingly, the information theoretic quantity that arises in our sample complexity analysis is Chernoff information and not Kullback-Leibler divergence. Chernoff information $d^*(x, y)$ between two Bernoulli distributions $\mathcal{B}(x)$ and $\mathcal{B}(y)$ is defined by

$$d^*(x, y) = d(z^*, x) = d(z^*, y) \text{ where } z^* \text{ is the unique } z \text{ such that } d(z, x) = d(z, y).$$

Our analysis of KL-Racing and KL-LUCB share the need to determine the number of samples of arm a needed to guarantee with high probability that some constant $c \in [p_{m+1}, p_m]$ does not belong to the interval $\mathcal{I}_a(t)$. Deriving such a result for intervals based on KL-divergence brings up Chernoff information :

Lemma 2
Let $T \geq 1$ *be an integer. Let* $\delta > 0$, $\gamma > 0$ *and* $c \in]0, 1[$ *such that* $p_a \neq c$.

$$\sum_{t=1}^{T} \mathbb{P}\left(a = u_t \vee a = l_t, N_a(t) > \left\lceil \frac{\gamma}{d^*(p_a, c)} \right\rceil, N_a(t) d(\hat{p}_a(t), c) \leq \gamma\right) \leq \frac{\exp(-\gamma)}{d^*(p_a, c)}.$$

Our analysis needs the sum of probabilities in Lemma 2 to be exponentially small in γ, which is the case when the number of samples of a is bigger than $\gamma / d^*(p_a, c)$. Cappé *et al.* (2013) bound the same probability in Appendix A.2, but for $N_a(t)$ bigger than $\gamma / d(p_a, c)$, and without getting the exponential decay we need.

3.2 Sample Complexity bounds

We state here the two sample complexity bounds we were able to obtain for KL-Racing and KL-LUCB. The latter involves the following complexity term depending on a parameter $c \in [p_{m+1}, p_m]$:

$$H^*_{\epsilon,c} := \sum_{a \in \{1,\dots,K\}} \frac{1}{\max(d^*(p_a, c), \epsilon^2/2)}$$

Theorem 1
Let $c \in [p_{m+1}, p_m]$. Let $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$, with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha - 1}$. The number of sample $\mathcal{N}$ used in KL-Racing with $\epsilon = 0$ is such that

$$\mathbb{P}\left(\mathcal{N} \leq \max_{a \in \{1,\dots,K\}} \frac{K}{d^*(p_a, c)} \log\left(\frac{k_1 K (H^*_{\epsilon,c})^\alpha}{\delta}\right) + 1, \mathcal{S}_\delta = \mathcal{S}^*_m\right) \geq 1 - 2\delta.$$

Theorem 2
Let $\epsilon \geq 0$. Let $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right) + \log\log\left(\frac{k_1 K t^\alpha}{\delta}\right)$. For $1 < \alpha \leq 1.2$ and $k_1 > 2e + 1 + \frac{e}{\alpha - 1} + \frac{e+1}{(\alpha-1)^2}$ KL-LUCB is δ-PAC and

$$\mathbb{P}\left(\mathcal{N} \leq 6 H^*_\epsilon \log\left(\frac{k_1 K (H^*_\epsilon)^{1.2}}{\delta}\right)\right) \geq 1 - 2\delta.$$

With $2 < \alpha \leq 2.2$ and $k_1 = 13$, KL-LUCB is δ-PAC and

$$\mathbb{E}[\mathcal{N}] \leq 24 H^*_\epsilon \log\left(\frac{13 (H^*_\epsilon)^{2.2}}{\delta}\right) + \frac{18\delta}{k_1(\alpha - 2)^2} \quad \text{with} \quad H^*_\epsilon = \min_{c \in [p_{m+1}; p_m]} H^*_{\epsilon,c}. \tag{4}$$

The result of Theorem 1 is only a first bound involving Chernoff information and might be improved to involve a sum over the different arms rather than a supremum. Moreover, this bound only holds on the event where the algorithm is correct, as also noted by Kalyanakrishnan *et al.* (2012) for successive elimination. For KL-LUCB, we are able to provide a more elegant result : an upper bound on the *expectation* of $\mathcal{N}$ involving the smaller quantity H^*_ϵ.

Although we have only considered Bernoulli distributions in this paper, note that KL-LUCB naturally extends to rewards in the exponential family (by using an appropriate d function, as shown by Cappé *et al.* (2013) for KL-UCB), and that Theorems 1 and 2 still hold in this more general setting.

4 Discussion on the complexity of Explore-m

We believe that the bound (4) on the expected sample complexity is the first of its kind involving KL-divergence (through Chernoff information). Pinsker's inequality shows that $d^*(x,y) \geq (x-y)^2/2$ and gives a relationship with the complexity term H_ϵ derived in Kalyanakrishnan *et al.* (2012) :

$$H_\epsilon = \sum_{a \in \{1,2,\dots K\}} \frac{1}{\max(\Delta_a^2, (\frac{\epsilon}{2})^2)}, \quad \text{with } \Delta_a = \begin{cases} p_a - p_{m+1} & \text{for } a \in \mathcal{S}^*_m, \\ p_m - p_a & \text{for } a \in (\mathcal{S}^*_m)^c. \end{cases}$$

On has $H^*_\epsilon \leq 8 H_\epsilon$. Although H^*_ϵ cannot be shown to be strictly smaller than H_ϵ on every problem, the explicit bound in (4) still improves over that of Kalyanakrishnan *et al.* (2012) in terms of the hidden constants. Also, the theoretical guarantees in Theorem 2 hold for smaller exploration rates, which appear to lower the sample complexity in practice. As the parameter c seems to be an avatar of the proof, we would rather conjecture the 'true' complexity term to be

$$\tilde{H}_\epsilon := \sum_{a \in \mathcal{S}^*_m} \frac{1}{\max(d^*(p_a, p_{m+1}), \frac{\epsilon^2}{2})} + \sum_{a \in (S^*_m)^c} \frac{1}{\max(d^*(p_a, p_m), \frac{\epsilon^2}{2})}. \tag{5}$$

Is there a fundamental difference between regret minimization and pure-exploration settings that would justify the introduction of Chernoff information in the latter ? Another reasonable guess if one expects the two settings to be more closely related would be

$$\overline{H}_\epsilon := \sum_{a \in \mathcal{S}^*_m} \frac{1}{\max(d(p_a, p_{m+1}), \frac{\epsilon^2}{2})} + \sum_{a \in (S^*_m)^c} \frac{1}{\max(d(p_a, p_m), \frac{\epsilon^2}{2})}. \tag{6}$$

A lower bound on the sample complexity of every algorithm that is δ-PAC is not currently known. Finding such a fundamental object would be an interesting direction for future work. Note that the existing lower bound for Explore-m of Kalyanakrishnan *et al.* (2012) is a worst-case result, stating that for every δ-PAC algorithm *there exists* a problem on which $\mathbb{E}[\mathcal{N}] \geq CH_\epsilon \log(m/\delta)$. This result should be improved in two directions : one would want a lower bound that holds for *every* bandit problem and involve the right information-theoretic quantity.

In pure-exploration problems, an alternative to Explore-m is the so-called 'fixed budget' setting, where a forecaster is asked to find the set of m best arms using a maximum number of samples, n, fixed in advance. In this setting, proposed by Audibert *et al.* (2010) for $m = 1$ (and $\epsilon = 0$) and generalized by Bubeck *et al.* (2013) to arbitrary values of m, the goal is to minimize the probability p_n of the recommendation to be wrong after the use of these n samples. A lower bound on the probability of error in the fixed budget setting is given by Audibert *et al.* (2010) for $m = 1$. Their result also involves the complexity term H_0 (for $\epsilon = 0$) introduced above and states that for any forecaster, on any bandit problem, $p_n \geq \exp(-C'n/H_0)$, for some constant C'. This result is no longer worst-case, but is not enlightening from an information-theoretic point of view. As explained by Gabillon *et al.* (2012), who propose an interesting comparison between Explore-m and Explore-m with fixed budget, these two settings appear to share the same complexity, at least when complexity is expressed in terms of gaps Δ_a. Knowing the information complexity of both settings (that might be different) would definitely be interesting.

5 Numerical experiments

On the basis of our theoretical analysis from the preceding sections, could we expect the "KL-ized" versions of our algorithms to perform better in practice ? Also, in our comparison between adaptive and uniform sampling algorithms, does our inability to provide a concrete expected-sample-complexity bound for the latter indicate a *practical* weakness ? In this section, we present numerical experiments that answer both these questions in the affirmative.

In our experiments, in addition to (KL-)LUCB and (KL-)Racing, we include (KL-)LSC, an adaptive sampling algorithm akin to (KL)-LUCB. This algorithm uses the same stopping criterion as (KL)-LUCB, but rather than sample arms u_t and l_t at stage t, (KL)-LSC samples the **l**east-**s**ampled arm from $J(t)$ (or $J(t)^c$) that **c**ollides (overlaps by at least ϵ) with some arm in $J(t)^c$ ($J(t)$). To ensure that all algorithms are provably PAC, we run them with the following parameters : (KL-)LUCB and (KL-)LSC with $\alpha = 1.1$, $k_1 = 405.5$, and (KL-)Racing with $\alpha = 1.1$, $k_1 = 11.1$. Results are summarized in Figure 1.

As a first order of business, we consider bandit instances with $K = 10, 20, \ldots, 60$ arms ; we generate 1000 random instances for each setting of K, with each arm's mean drawn uniformly at random from $[0, 1]$. We set $m = \frac{K}{5}$, $\epsilon = 0.1$, $\delta = 0.1$. The expected sample complexity of each algorithm on the bandit instances for each K are plotted in Figure 5. Indeed we observe for each K that (1) the KL-ized version of each algorithm enjoys a lower sample complexity, and (2) (KL)-LUCB outperforms (KL-)LSC, which outperforms (KL-)Racing. The differences in sample complexity consistently increase with K.

These trends, aggregated from multiple bandit instances, indeed hold for nearly every individual bandit instance therein. In fact, we find that KL-izing has a more pronounced effect on bandit instances with means close to 0 or 1. For illustration, consider instance B_1 ($K = 15$; $p_1 = \frac{1}{2}$; $p_a = \frac{1}{2} - \frac{a}{40}$ for $a = 2, 3, \ldots, K$), an instance used by (Bubeck *et al.*, 2013, see Experiment 5). Figure 5 compares the runs of LUCB and KL-LUCB both on B_1 (with $m = 3$, $\epsilon = 0.04$, $\delta = 0.1$), and a "scaled-down" version B_2 (with $m = 3$, $\epsilon = 0.02$, $\delta = 0.1$) in which each arm's mean is half that of the corresponding arm's in B_1 (and thus closer to 0). While LUCB and KL-LUCB both incur a higher sample complexity on the harder B_2, the latter's relative economy is clearly visible in the graph—an advantage that could benefit applications such as optimizing click-through rates of on-line advertisements.

How conservative are the stopping criteria of our PAC algorithms ? In our third experiment, we halt these algorithms at intervals of 1000 samples, and at each stage record the probability that the set $J(t)$ of m empirical best arms that would be returned at that stage is non-optimal. Results from this experiment, again on B_1 , are plotted in Figure 5. Notice that (KL)-LUCB indeed drives down the mistake probability much faster than its competitors. Yet, even if all the algorithms have an empirical mistake probability smaller than δ after 5,000 samples, they only stop after at least 20,000 episodes, leaving us to conclude that our formal bounds are rather conservative.

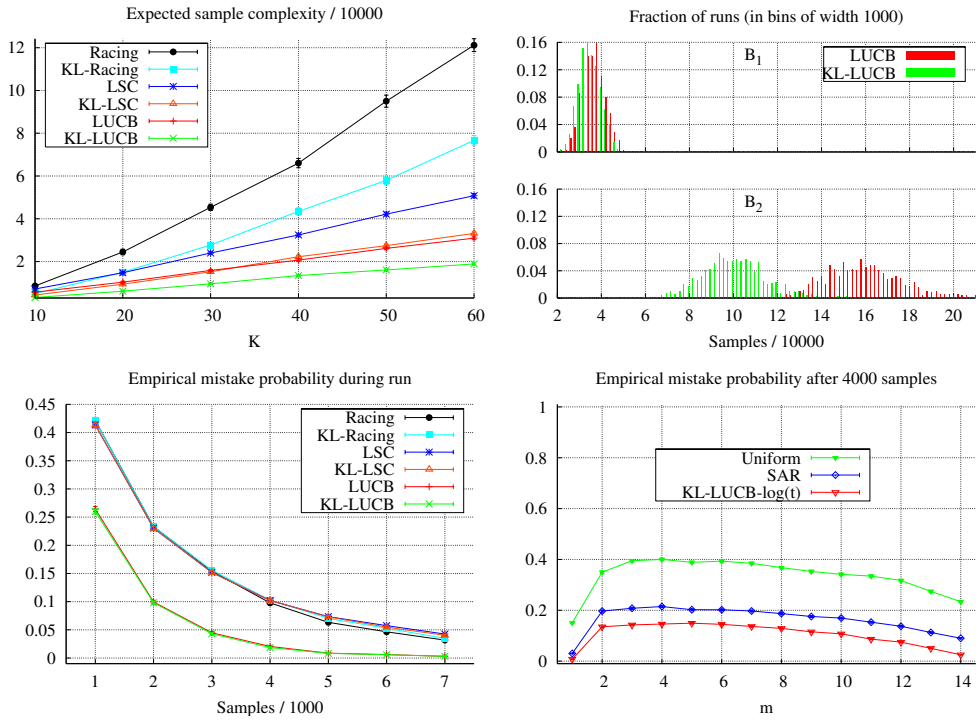

FIGURE 1 – Experimental results (descriptions provided in text).

In response, we test KL-LUCB-$log(t)$, a version of KL-LUCB with an exploration rate of $log(t)$ (which yields no provable guarantees) as a candidate for *Explore-m with fixed budget*, a setting described at the end of Section 4. Existing algorithms for this setting are mostly based on uniform sampling ideas. The SAR algorithm (for Successive Accepts and Rejects) of Bubeck *et al.* (2013) sets some 'deadlines' in advance (depending on the budget n) at which one arm exactly should be selected or discarded and samples uniformly the arms between two deadlines. As seen on Figure 5, that displays an empirical estimate of p_n (the probability of error) as a function of the number of arms m to find, KL-LUCB-$log(t)$ appears to outperforms the SAR algorithm of Bubeck *et al.* (2013), showing adaptive sampling might be a good idea in the fixed budget setting too. As future work, it would be very relevant to consider formal error bounds for adaptive sampling algorithms such as KL-LUCB-$log(t)$ in this alternative setting.

6 Conclusion

This paper presents a successful translation of recent improvements for bandit problems in the regret setting to the pure-exploration setting. Incorporating confidence intervals based on KL-divergence into the uniform and adaptive sampling heuristics, which have been used previously for Explore-m, we introduce the KL-LUCB and KL-Racing algorithms, which improve both in theory and in practice over their Hoeffding counterparts. Our experiments also provide the novel insight that adaptive sampling might be superior to uniform sampling and eliminations.

For KL-LUCB, we provide the first finite-time upper bound on the expected sample complexity involving Chernoff information. Is there a fundamental difference between the regret and pure-exploration settings that would justify a different complexity measure, albeit one still based on KL-divergence ? A problem-dependent lower bound on the expected sample complexity of any PAC algorithm for Explore-m could answer this question, and is left as an interesting open question. As another gap between regret and pure-exploration, one might consider that no counterpart of the Thompson Sampling algorithm, recently shown to be optimal in the regret setting Kaufmann *et al.* (2012) as well as practically very efficient, has yet been found for Explore-m.

Références

AUDIBERT J.-Y., BUBECK S. & MUNOS R. (2010). Best arm identification in multi-armed bandits. In *Conference on Learning Theory (COLT)*.

AUER P., CESA-BIANCHI N. & FISCHER P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, **47**(2), 235–256.

BUBECK S., MUNOS R. & STOLTZ G. (2011). Pure exploration in finitely armed and continuous armed bandits. *Theoretical Computer Science 412, 1832-1852*, **412**, 1832–1852.

BUBECK S., WANG T. & VISWANATHAN N. (2013). Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning (ICML)* : To appear.

CAPPÉ O., GARIVIER A., MAILLARD O.-A., MUNOS R. & STOLTZ G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *to appear in Annals of Statistics*.

EVEN-DAR E., MANNOR S. & MANSOUR Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, **7**, 1079–1105.

GABILLON V., GHAVAMZADEH M. & LAZARIC A. (2012). Best arm identification : A unified approach to fixed budget and fixed confidence. In *Neural Information and Signal Processing (NIPS)*.

GARIVIER A. & CAPPÉ O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory (COLT)*.

HEIDRICH-MEISNER V. & IGEL C. (2009). Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *International Conference on Learning Theorey (ICML)*.

KALYANAKRISHNAN S. & STONE P. (2010). Efficient selection in multiple bandit arms : Theory and practice. In *International Conference on Machine Learning (ICML)*.

KALYANAKRISHNAN S., TEWARI A., AUER P. & STONE P. (2012). PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

KAUFMANN E., KORDA N. & MUNOS R. (2012). Thompson sampling : an asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory (ALT)*.

LAI T. & ROBBINS H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**(1), 4–22.

MAILLARD O.-A., MUNOS R. & STOLTZ G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Conference On Learning Theory (COLT)*.

MARON O. & MOORE A. (1997). The racing algorithm : Model selection for lazy learners. *Artificial Intelligence Review*, **11(1-5)**, 113–131.

MNIH V., SZEPESVÁRI C. & AUDIBERT J.-Y. (2008). Empirical Bernstein stopping. In *International Conference on Machine Learning (ICML)*.

THOMPSON W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.