Information complexity in bandit subset selection

Emilie Kaufmann & Shivaram Kalyanakrishnan

In a nutshell

We present **two improved algorithms for Explore-m** based on different heuristics, adaptive and uniform sampling, and sharing the use of confidence intervals based on KL-divergence. A new information-theoretic quantity, 'Chernoff information', arises in their analysis, while their practical performance give the new insight that adaptive sampling might be superior to uniform sampling.

The Explore-m problem

A bandit model is a collection of K arms. Arm a is a unknown Bernoulli distribution $\mathcal{B}(p_a)$. Drawing arm a is observing a sample from $\mathcal{B}(p_a)$. Assume $p_1 \geq \dots p_m \geq p_{m+1} \geq \dots p_K$. The set of (ϵ, m)-optimal arms is

$$\mathcal{S}^*_{m,\epsilon} = \{a : p_a \geq p_m + \epsilon\}.$$

A forecaster interacting with a bandit model:

- adopts a sampling strategy to decide which arm to draw at which round
- stops playing after observing a (possibly random) number of samples $\mathcal{N}$ from the arms and recommends a set $\mathcal{S}$ of m arms

Two **pure-exploration problems**: find an algorithm that:

Explore-m:

- satisfies $\mathbb{P}(\mathcal{S} \subset \mathcal{S}^*_{m,\epsilon}) \geq 1-\delta$ (δ-PAC algorithm)
- minimizes the expected *sample complexity* $\mathbb{E}[\mathcal{N}]$.

Explore-m with fixed budget:

- satisfies $\mathcal{N} \leq n$, where n is a *budget known in advance.*
- minimizes the probability of error $p_n := \mathbb{P}(\mathcal{S} \subset \mathcal{S}^*_{m,\epsilon})$

⚠ These two problems are different from regret minimization in bandit models.

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n}(p_1 - p_{A_t})\right] \text{ when } A_t \text{ is the arm drawn at time } t$$

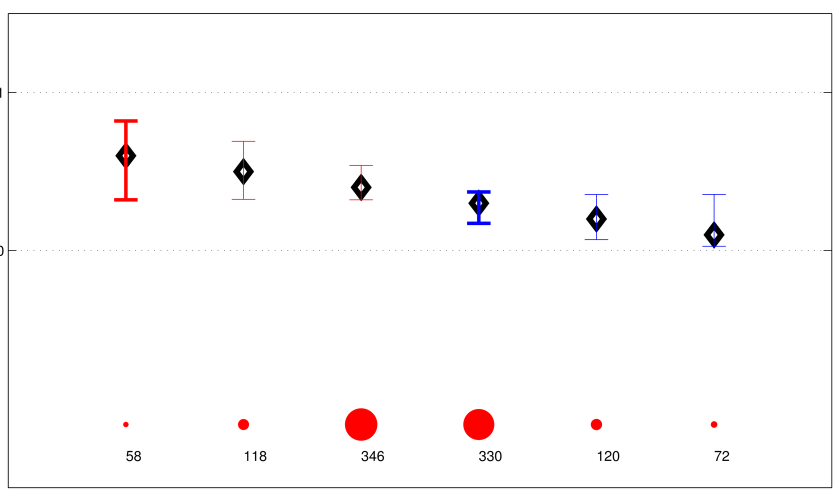
The KL-LUCB algorithm

Let $J(t)$ be the m arms with the highest empirical means at time t and u_t and l_t, two 'critical' arms likely to be misclassified:

$$u_t = \text{argmax}_{j \notin J(t)} U_j(t) \quad \text{and} \quad l_t = \text{argmin}_{j \in J(t)} L_j(t). \qquad (1)$$

At each round KL-LUCB:

- Samples two arms adaptively chosen from the past, arms u_t and l_t
- Stops when $U_{u_t} - L_{l_t} < \epsilon$ and recommends $J(t)$.



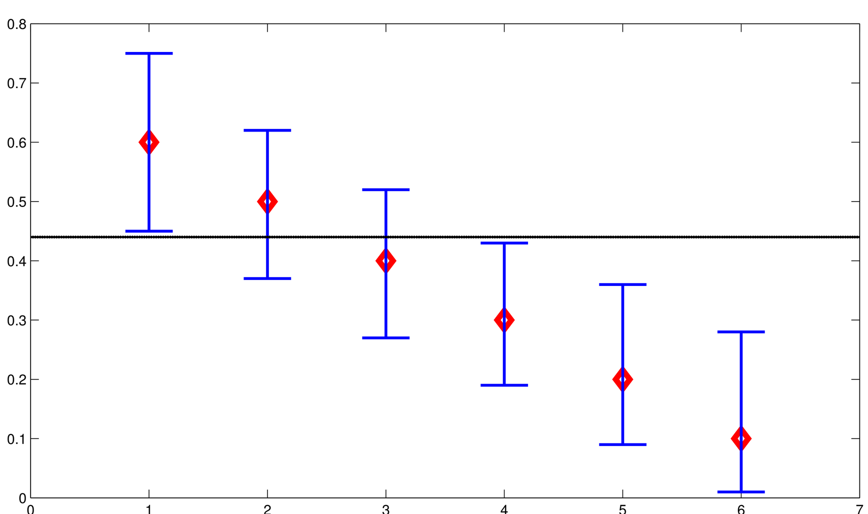
A stopping configuration for $m = 3$. Arms from $J(t)$ are separated from $J(t)^c$ by $\epsilon = 0.05$

The KL-Racing algorithm

$\mathcal{R}$ be the set of remaining arms, $\mathcal{S}$ of selected arm and $\mathcal{D}$ of discarded arm.

At round t, KL-Racing:

- samples all the arms in $\mathcal{R}$ (i.e. samples uniformly the remaining arms)
- compute $J(t)$ the set of $m - |\mathcal{S}|$ empirical best arms, $J(t)^c = \mathcal{R} \setminus J(t)$
- selects the empirical best arm of $\mathcal{R}$, a_B if $L_{a_B}(t) > U_{u_t}(t) - \epsilon$
  $$\mathcal{S} = \mathcal{S} \cup \{a_B\}, \quad \mathcal{R} = \mathcal{R} \setminus \{a_B\}$$
- discards the empirical worst arm of $\mathcal{R}$, a_W if $U_{a_W}(t) < L_{l_t}(t) + \epsilon$
  $$\mathcal{D} = \mathcal{D} \cup \{a_W\}, \quad \mathcal{R} = \mathcal{R} \setminus \{a_W\}$$



The optimal arm is selected, since its LCB is bigger than the UCB of the $K - m$ worst arms ($m = 3$)

What is the complexity of Explore-m?

The regret minimization problem is 'solved' since for any sampling strategy

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq \sum_{a=2}^{K} \frac{p_1 - p_a}{d(p_a, p_1)} \quad \text{with } d(p,p') := \text{KL}(\mathcal{B}(p), \mathcal{B}(p'))$$

and there exists algorithms matching this lower bound (e.g. KL-UCB).

For Explore-m, upper bounds on $\mathbb{E}[\mathcal{N}]$ for some δ-PAC algorithms scale in $O\left(H_\epsilon \log\left(\frac{H_\epsilon}{\delta}\right)\right)$, where

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

And a lower bound on $\mathbb{E}[\mathcal{N}]$ for every δ-PAC algorithm is not currently known.

The 'true' complexity of Explore-m must involve some information theoretic quantity . Is it Kullback Leibler divergence d or Chernoff information d^* ?

$$d^*(p,p') := d(p^*, p) = d(p^*, p') \text{ where } p^* \text{ is defined by } d(p^*, p) = d(p^*, p')$$

Upper bounds on the sample complexity of the algorithms we propose involve

$$H^*_{\epsilon,c} := \sum_{a \in \{1,2,\dots,K\}} \frac{1}{\max(d^*(p_a, c), \epsilon^2/2)} \quad \text{for } c \in [p_{m+1}, p_m]$$

Algorithms: two heuristics

Existing algorithms broadly fall into two categories:

- uniform sampling and eliminations (Racing)
- adaptive sampling (LUCB)

Racing and LUCB are two generic algorithms based on confidence intervals for the parameter of each arm, $\mathcal{I}_a(t) = [L_a(t), U_a(t)]$. We analyze the version of these algorithm using confidence intervals based on KL-divergence :

$$\begin{aligned} U_a(t) &= u_a(t) &:=& \max\{q \in [\hat{p}_a(t), 1] : N_a(t) d(\hat{p}_a(t), q) \leq \beta(t,\delta)\}, \text{ and} \\ L_a(t) &= l_a(t) &:=& \min\{q \in [0, \hat{p}_a(t)] : N_a(t) d(\hat{p}_a(t), q) \leq \beta(t,\delta)\}, \end{aligned}$$

for some exploration rate $\beta(t,\delta)$.

Theoretical guarantees

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 $k_1 > 1 + \frac{1}{\alpha - 1}$. KL-Racing with $\epsilon = 0$ is δ-PAC and the number of samples $\mathcal{N}$ satisfies:*

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

Theorem 2 *Let $\epsilon \geq 0$. Let $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right) + \log\log\left(\frac{k_1 K t^\alpha}{\delta}\right)$. With $2 < \alpha \leq 2.2$ and $k_1 = 13$, KL-LUCB is δ-PAC and*

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

- Conjecture on a lower bound for the sample complexity

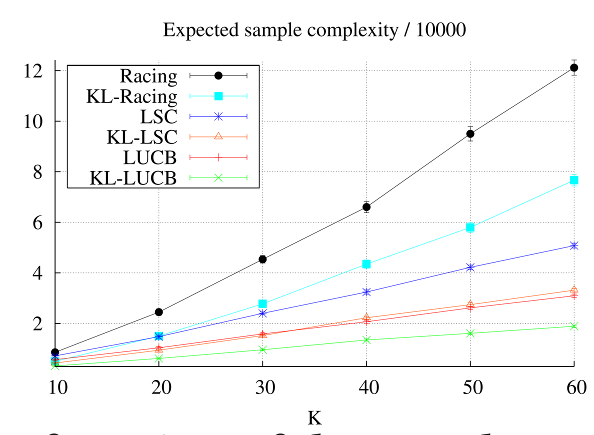
$$\mathbb{E}[\mathcal{N}] \geq \left(\sum_{a \in \mathcal{S}^*_m} \frac{1}{\max(d^*(p_a, p_{m+1}), \frac{\epsilon^2}{2})} + \sum_{a \in (\mathcal{S}^*_m)^c} \frac{1}{\max(d^*(p_a, p_m), \frac{\epsilon^2}{2})}\right) \log\left(\frac{1}{\delta}\right)$$

or

$$\mathbb{E}[\mathcal{N}] \geq \left(\sum_{a \in \mathcal{S}^*_m} \frac{1}{\max(d(p_a, p_{m+1}), \frac{\epsilon^2}{2})} + \sum_{a \in (\mathcal{S}^*_m)^c} \frac{1}{\max(d(p_a, p_m), \frac{\epsilon^2}{2})}\right) \log\left(\frac{1}{\delta}\right) ?$$

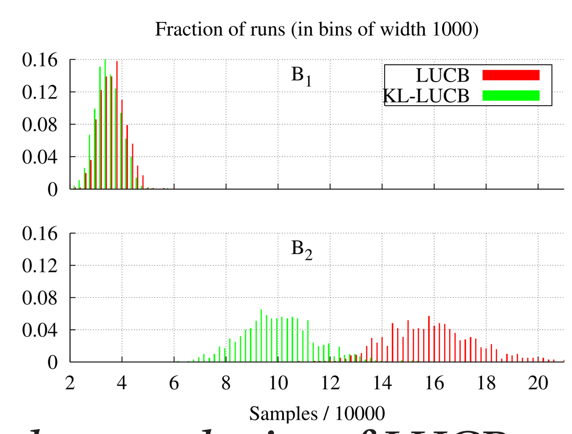
Practical performance

- Adaptive sampling seems superior to uniform sampling and eliminations



Sample complexity as a function of the number of arms K in the problem (setting $m = K/5$), averaged over 1000 problems picked uniformly at random

- Using KL-based confidence intervals drifts down the sample complexity



Distribution of the sample complexity of LUCB and KL-LUCB on two fixed problems, $B_1 : K = 15$; $p_1 = \frac{1}{2}$; $p_a = \frac{1}{2} - \frac{a}{40}$ for $a = 2, 3, \dots, K$ and $B_2 = \frac{1}{2} B_1$