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 gives 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 known Bernoulli distribution \(B(p_a)\). Drawing arm \(a\) is observing a sample from \(B(p_a)\). Assume \(p_1 \geq \ldots \geq p_m \geq p_{m+1} \geq \ldots \geq p_K\). The set of \((c, m)\)-optimal arms is

\[ S_{m,c} = \{a: p_a \geq p_m + c\} \]

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 \(N\) from the arms and recommends a set \(S\) of \(m\) arms

Two pure-exploration problems: find an algorithm that:

<table>
<thead>
<tr>
<th>Explore-(m)</th>
<th>Explore-(m) with fixed budget:</th>
</tr>
</thead>
<tbody>
<tr>
<td>satisfies (P(\mathcal{F} \subset \mathcal{S}_{m,c}) \geq 1 - \delta) ((\delta)-PAC algorithm)</td>
<td>satisfies (N \leq n), where (n) is a budget-given in advance.</td>
</tr>
<tr>
<td>minimizes the expected sample complexity (E[N])</td>
<td>minimizes the probability of error (p_m := P(\mathcal{F} \subset \mathcal{S}_{m,c})).</td>
</tr>
</tbody>
</table>

These two problems are different from regret minimization in bandit models:

- for each round KL-LUCB:
  \[ R_m = \mathbb{E}(\sum_{t=1}^{\infty} (p_t - p_\ast)) \text{ when } A_t \text{ is the arm drawn at time } t \]
- the optimal arm is selected, since its LCB is bigger than the UCB of the \(K\) worst arms \((m = 3)\)

What is the complexity of Explore-\(m\)?

The regret minimization problem is ‘solved’ since for any sampling strategy

\[ \liminf_{n \to \infty} \frac{R_n}{\log(n)} \leq K \frac{p_\ast - p_m}{\log K} \]

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

For Explore-\(m\), upper bounds on \(E[N]\) for some \(\delta\)-PAC algorithms scale in \(O(\sqrt{KL \log(\frac{1}{\delta})})\), where

\[ H_\delta = \sum_{a \in \{1, \ldots, K\}} \max \{\frac{2}{\delta} \}, \quad \Delta_a = \begin{cases} p_a - p_{m+1} & \text{for } a \in S_{m,c}^\ast, \\ p_m - p_a & \text{for } a \in \mathcal{S}_{m,c}^\ast. \end{cases} \]

And a lower bound on \(E[N]\) for every \(\delta\)-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 \(d'\)?

\[ d^\ast(p, p') := d(p\ast, p') = d(p', p\ast) \]

where \(p\ast\) is defined by \(d(p\ast, p') = d(p', p\ast)\).

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

[\(H^\ast_\delta := \sum_{a \in \{1, \ldots, K\}} \max \{d^\ast(p_a, c), e^2/2\}\) 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 (UCB)

Racing and LUCB are two generic algorithms based on confidence intervals for the parameter of each arm, \(\mathcal{J}_t(a) = [L_t(a), U_t(a)]\). We analyze the version of these algorithm using confidence intervals based on KL-divergence:

\[ U_t(a) = u_t(a) := \max \{q \in [p(a, t)(1), N_t(a)(d(p_a, q), \delta) \leq \beta(\delta, \delta)\}, \quad L_t(a) := l_t(a) := \min \{q \in [p(a, t)(1), N_t(a)(d(p_a, q), \delta) \leq \beta(\delta, \delta)\}, \]

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

THEORETICAL GUARANTEES

**Theorem 1** Let \(c \in (p_{m+1}, p_m]\). Let \(\beta(\delta, \delta) = \log \left(\frac{1 + e^2}{\delta}\right)\), with \(a > 1\) and \(k_1 > 1 + \frac{1}{a}\). KL-Racing with \(c = 0\) is \(\delta\)-PAC and the number of samples \(N\) satisfies:

\[ P(\mathcal{J}_t(a) \cap \mathcal{S}_{m,c}^\ast) \leq K \frac{p_\ast - p_m}{\log K} \left(1 + \min_{a \in \{p_{m+1}, p_m\}} \frac{1}{\beta(\delta, \delta)} H^\ast_\delta\right) \leq K (1 - 2\delta) \]

**Theorem 2** Let \(c = 0\). Let \(\beta(\delta, \delta) = \log \left(\frac{1 + e^2}{\delta}\right) + \log \frac{1}{a}\). With \(2 < a \leq 2\)

\[ \mathcal{E}(N) \leq 24H^\ast_\delta \log \frac{13(1/\delta)^{1/2}}{b} + \frac{185}{10} \log \frac{1}{b} \] with \(H^\ast_\delta = \min_{c \in (p_{m+1}, p_m]} H^\ast_\delta\).

Conjecture on a lower bound for the sample complexity

\[ \mathcal{E}(N) \geq \sum_{a \in \mathcal{S}_{m,c}^\ast} \max \{d^\ast(p_a, c), e^2/2\} \]

or

\[ \mathcal{E}(N) \geq \sum_{a \in \mathcal{S}_{m,c}^\ast} \max \{d^\ast(p_a, c), e^2/2\} \log \left(\frac{1}{\delta}\right) \]

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