Information complexity in bandit subset selection

Emilie Kaufmann & Shivaram Kalyanakrishan

Telecom ParisTech, Yahoo! Labs, Bangalore

COLT 2013, Princeton University

Stochastic Bandit: One statistical model

A **multi-armed bandit model** is a set of K arms where

- Each arm a is a Bernoulli distribution of unknown parameter p_a
- Drawing arm a is observing a realization of $\mathcal{B}(p_a)$

At round t, a forecaster

- chooses arm A_t to draw based on past observations, according to its **sampling strategy**
- observes 'reward' $X_t \sim \mathcal{B}(p_{A_t})$

Its goal is to **learn which arm is the best**:

$$a^* = \text{argmax}_a \; p_a$$

Stochastic Bandit: Two objectives

- Regret minimization

The forecaster wants to **maximize the reward accumulated during learning** or equivalentely minimize its **regret**:

$$R_n = np_{a^*} - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

$\Rightarrow$ his sampling strategy **tradeoffs exploration and exploitation**

Stochastic Bandit: Two objectives

- Regret minimization

The forecaster wants to **maximize the reward accumulated during learning** or equivalentely minimize its **regret**:

$$R_n = np_{a^*} - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$$

$\Rightarrow$ his sampling strategy **tradeoffs exploration and exploitation**

- Pure exploration

The forecaster has to **recommend the (set of) best(s) arm(s)** using as few observations of the arms as possible. (no loss for 'bad' observations)

$\Rightarrow$ his sampling strategy **optimaly explores the environnement**

The Explore-m problem

Assume $p_1 \geq \cdots \geq p_m \geq p_{m+1} \geq \ldots p_K$.

Parameters and notations

- m be the number of arms to find
- $\delta \in]0,1[$ confidence parameter, $\epsilon \geq 0$ tolerance parameter
- $\mathcal{S}^*_{m,\epsilon} = \{a : p_a \geq p_m - \epsilon\}$ be the set of (ϵ, m)-optimal arms

The Explore-m problem

Assume $p_1 \geq \cdots \geq p_m \geq p_{m+1} \geq \ldots p_K$.

Parameters and notations

- m be the number of arms to find
- $\delta \in]0,1[$ confidence parameter, $\epsilon \geq 0$ tolerance parameter
- $\mathcal{S}^*_{m,\epsilon} = \{a : p_a \geq p_m - \epsilon\}$ be the set of (ϵ, m)-optimal arms

The forecaster

- chooses at time t one (or several) arms to draw
- decides to stop after a (possibly random) total number of samples from the arms $\mathcal{N}$
- recommends a set $\mathcal{S}$ of m arms

The Explore-m problem

Assume $p_1 \geq \cdots \geq p_m \geq p_{m+1} \geq \ldots p_K$.

Parameters and notations

- m be the number of arms to find
- $\delta \in]0,1[$ confidence parameter, $\epsilon \geq 0$ tolerance parameter
- $\mathcal{S}^*_{m,\epsilon} = \{a : p_a \geq p_m - \epsilon\}$ be the set of (ϵ, m)-optimal arms

The forecaster

- chooses at time t one (or several) arms to draw
- decides to stop after a (possibly random) total number of samples from the arms $\mathcal{N}$
- recommends a set $\mathcal{S}$ of m arms

His goal

- $\mathbb{P}(\mathcal{S} \subseteq \mathcal{S}^*_{m,\epsilon}) \geq 1-\delta$, and $\mathbb{E}[\mathcal{N}]$ is small

Open questions for Explore-m

The regret minimization problem is 'solved' in some sense:

- An (asymptotic) lower bound on the regret of any good algorithm

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{\mathsf{KL}(\mathcal{B}(p_a), \mathcal{B}(p_1))}$$

- Algorithms matching this lower bound: KL-UCB, Thompson Sampling

Open questions for Explore-m

The regret minimization problem is 'solved' in some sense:

- An (asymptotic) lower bound on the regret of any good algorithm

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{\mathsf{KL}(\mathcal{B}(p_a), \mathcal{B}(p_1))}$$

- Algorithms matching this lower bound: KL-UCB, Thompson Sampling

Is there also an 'information' complexity for Explore-m?

- For any algorithm correct with probability $1-\delta$,

$$\mathbb{E}[\mathcal{N}] \geq \ ?$$

Our contributions

- Algorithms for Explore-m broadly fall into two categories: **uniform sampling and eliminations** versus **adaptive sampling**, both using **confidence intervals**

Our contributions

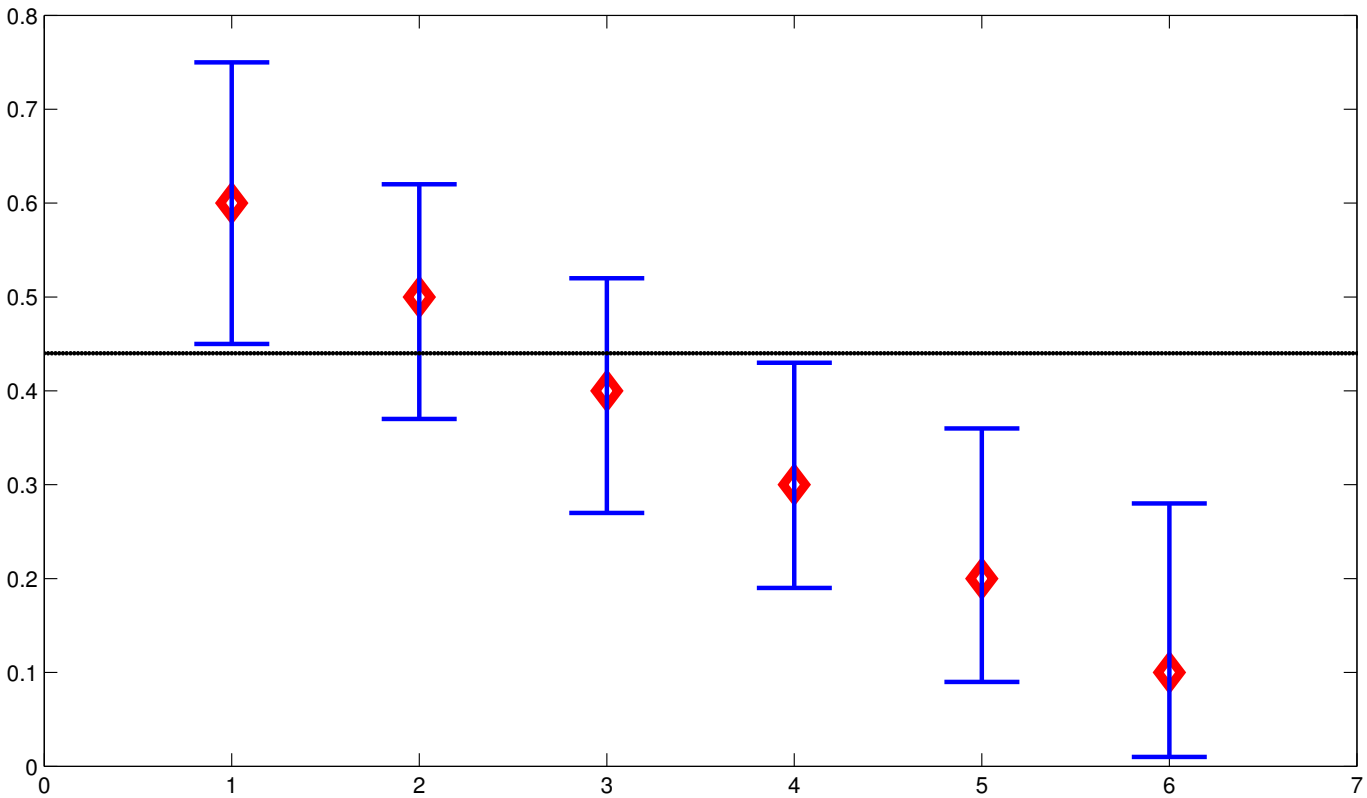
- Algorithms for Explore-m broadly fall into two categories: **uniform sampling and eliminations** versus **adaptive sampling**, both using **confidence intervals**
- We transpose the use of **improved KL-based confidence intervals** from the regret minimization to the pure-exploration setting

Our contributions

- Algorithms for Explore-m broadly fall into two categories: **uniform sampling and eliminations** versus **adaptive sampling**, both using **confidence intervals**
- We transpose the use of **improved KL-based confidence intervals** from the regret minimization to the pure-exploration setting
- Upper bound on the complexity of our algorithms involve KL-divergence through a quantity named **Chernoff information**

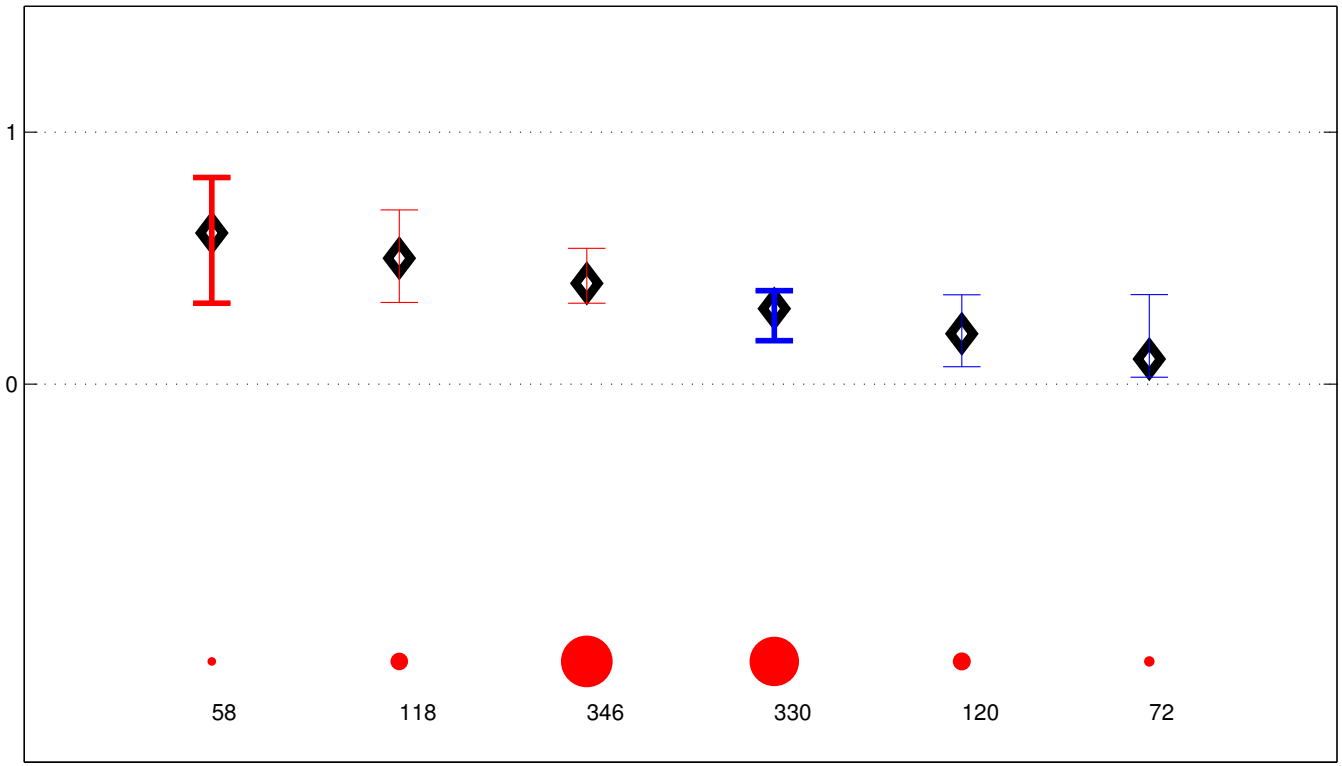
KL-Racing: uniform sampling and eliminations

$$p = [0.6\ 0.5\ 0.4\ 0.3\ 0.2\ 0.1] \quad m = 3 \quad \delta = 0.1 \quad \epsilon = 0$$



In this situation, arm 1 is selected (elimination)

KL-LUCB: adaptive sampling



Arms in bold arms are sampled. The stopping condition is matched.

Sample Complexity bound for KL-LUCB

For KL-LUCB, we have that $\mathbb{E}[\mathcal{N}] = O\left(H_\epsilon^* \log\left(\frac{1}{\delta}\right)\right)$ with

$$H_\epsilon^* = \min_{c \in [p_{m+1}; p_m]} \sum_{a=1}^{K} \frac{1}{\max(K^*(p_a, c), \epsilon^2/2)}.$$

where K^* is the Chernoff information:

$$K^*(x, y) := K(z^*, x) = K(z^*, y),$$