Best Arm Identification in Multi-Armed Bandits

Jean-Yves Audibert[1,2] & Sébastien Bubeck[3] & Rémi Munos[3]

[1] Univ. Paris Est, Imagine
[2] CNRS/ENS/INRIA, Willow project
[3] INRIA Lille, SequeL team

Best arm identification task

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions (over $[0,1]$) $\nu_1, \ldots, \nu_K$ of the arms. We assume that there is a unique arm i^* with maximal mean.

For each round $t = 1, 2, \ldots, n$;

1. The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$.
2. The environment draws the reward Y_t from ν_{I_t} (and independently from the past given I_t).

At the end of the n rounds the forecaster outputs a recommendation $J_n \in \{1, \ldots, K\}$.

Goal: Find the best arm, i.e, the arm with maximal mean. Regret:

$$e_n = \mathbb{P}(J_n \neq i^*).$$

Motivating examples

- Clinical trials for cosmetic products. During the **test phase**, several several formulæ for a cream are **sequentially tested**, and after a finite time **one is chosen** for commercialization.
- Channel allocation for mobile phone communications. Cellphones can **explore the set of channels** to find the best one to operate. Each **evaluation** of a channel is **noisy** and there is a **limited number** of evaluations before the communication starts on **the chosen channel**.

Summary of the talk

- Let μ_i be the mean of ν_i, and $\Delta_i = \mu_{i^*} - \mu_i$ the suboptimality of arm i.
- Main theoretical result: it requires of order of $H = \sum_{i \neq i^*} 1/\Delta_i^2$ rounds to find the best arm. Note that this result is well known for $K = 2$.
- We present two new forecasters, **Successive Rejects (SR)** and **Adaptive UCB-E (Upper Confidence Bound Exploration)**.
- SR is parameter free, and has optimal guarantees (up to a logarithmic factor).
- Adaptive UCB-E has no theoretical guarantees but it experimentally outperforms SR.

Lower Bound

Theorem

Let $\nu_1, \ldots, \nu_K$ be Bernoulli distributions with parameters in $[1/3, 2/3]$. There exists a numerical constant $c > 0$ such that for any forecaster, up to a permutation of the arms,

$$e_n \geq \exp\left(-c(1 + o(1))\frac{n \log(K)}{H}\right).$$

Informally, any algorithm requires at least (of order of) $H/\log(K)$ rounds to find the best arm.

Lower Bound

Theorem

Let $\nu_1, \ldots, \nu_K$ be Bernoulli distributions with parameters in $[1/3, 2/3]$. There exists a numerical constant $c > 0$ such that for any forecaster, up to a permutation of the arms,

$$e_n \geq \exp\left(-c\left(1 + \frac{K\log(K)}{\sqrt{n}}\right)\frac{n\log(K)}{H}\right).$$

Informally, any algorithm requires at least (of order of) $H/\log(K)$ rounds to find the best arm.

Uniform strategy

For each $i \in \{1, \dots, K\}$, select arm i during $\lfloor n/K \rfloor$ rounds. Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,\lfloor n/K \rfloor}$.

Theorem

The uniform strategy satisfies: $e_n \leq 2K \exp\left(-\frac{n \min_i \Delta_i^2}{2K}\right)$.
For any $(\delta_1, \dots, \delta_K)$ *with* $\min_i \delta_i \leq 1/2$*, there exist distributions such that* $\Delta_1 = \delta_1, \dots, \Delta_K = \delta_K$ *and*

$$e_n \geq \frac{1}{2} \exp\left(-\frac{8n \min_i \Delta_i^2}{K}\right).$$

Informally, the uniform strategy finds the best arm with (of order of) $K / \min_i \Delta_i^2$ rounds. For large K, this can be significantly larger than $H = \sum_{i \neq i^*} 1/\Delta_i^2$.

UCB-E

Draw each arm once

For each round $t = K + 1, 2, \dots, n$:

$$\text{Draw } I_t \in \operatorname*{argmax}_{i \in \{1,\dots,K\}} \left(\widehat{X}_{i,T_i(t-1)} + \sqrt{\frac{n/H}{2T_i(t-1)}} \right),$$

where $T_i(t-1)$ = nb of times we pulled arm i up to time $t-1$.

Let $J_n \in \operatorname{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Theorem

UCB-E satisfies $e_n \leq n \exp\left(-\frac{n}{50H}\right)$.

UCB-E finds the best arm with (of order of) H rounds, but it requires the knowledge of $H = \sum_{i \neq i^*} 1/\Delta_i^2$.

Successive Rejects (SR)

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \dots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \dots, K-1\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Motivation for choosing n_k

Consider $\mu_1 > \mu_2 = \dots = \mu_M \gg \mu_{M+1} = \dots = \mu_K$

- target: draw n/M times the M best arms
- SR: the M best arms are drawn more than $n_{K-M+1} \approx \frac{1}{\overline{\log}(K)} \frac{n}{M}$

Successive Rejects (SR)

Let $\overline{\log(K)} = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $A_1 = \{1, \dots, K\}$, $n_0 = 0$ and $n_k = \lceil \frac{1}{\overline{\log(K)}} \frac{n-K}{K+1-k} \rceil$ for $k \in \{1, \dots, K-1\}$.

For each phase $k = 1, 2, \dots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{X}_{i,n_k}$, where $\widehat{X}_{i,s}$ represents the empirical mean of arm i after s pulls.

Let J_n be the unique element of A_K.

Theorem

SR satisfies:

$$e_n \leq K \exp\left(-\frac{n}{4H \log K}\right).$$

UCB-E

Parameter: exploration constant $c > 0$.

Draw each arm once

For each round $t = 1, 2, \dots, n$:

$$\text{Draw } I_t \in \operatorname*{argmax}_{i \in \{1,\dots,K\}} \left(\widehat{X}_{i,T_i(t-1)} + \sqrt{\frac{c\ n/H}{T_i(t-1)}} \right),$$

where $T_i(t-1) =$ nb of times we pulled arm i up to time $t-1$.

Let $J_n \in \operatorname{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

Adaptive UCB-E

Parameter: exploration constant $c > 0$.

For each round $t = 1, 2, \dots, n$:

(1) Compute an (under)estimate $\hat{H}_t$ of H

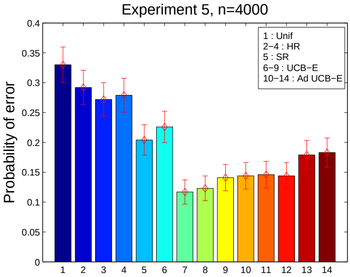
(2) Draw $I_t \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \left(\widehat{X}_{i,T_i(t-1)} + \sqrt{\frac{c\ n/\hat{H}_t}{T_i(t-1)}} \right)$,

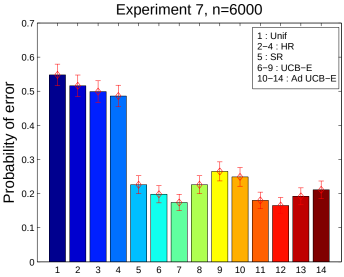
Let $J_n \in \mathrm{argmax}_{i \in \{1,\dots,K\}} \widehat{X}_{i,T_i(n)}$.

- Overestimating H $\Rightarrow$ low exploration of the arms $\Rightarrow$ potential missing of the optimal arm $\Rightarrow$ all Δ_i badly estimated
- Underestimating H $\Rightarrow$ higher exploration $\Rightarrow$ not focusing enough on the arms $\Rightarrow$ bad estimation of $H = \sum_{i \neq i^*} 1/\Delta_i^2$

Experiments with Bernoulli distributions

- Experiment 5: Arithmetic progression, $K = 15$, $\mu_i = 0.5 - 0.025i$, $i \in \{1, \dots, 15\}$.
- Experiment 7: Three groups of bad arms, $K = 30$, $\mu_1 = 0.5$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$.





Conclusion

- It requires at least $H/\log(K)$ rounds to find the best arm, with $H = \sum_{i \neq i^*} 1/\Delta_i^2$.
- UCB-E requires only $H \log n$ rounds but also the knowledge of H to tune its parameter.
- SR is a parameter free algorithm that requires less than $H \log^2 K$ rounds to find the best arm.
- Adaptive UCB-E does not have theoretical guarantees but it experimentally outperforms SR.