Profitable Bandits

Mastane Achab, Stephan Clémençon,
Aurélien Garivier

ACML 2018, Beijing Jiaotong University,
Beijing, China

Outline

Outline

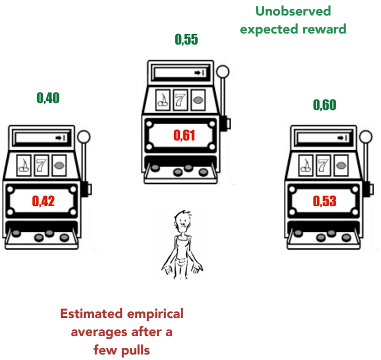
The classical Multi-Armed Bandit problem

At each time $t = 1, ..., T$

- Pull an arm $a_t \in \{1, \ldots, K\}$
- Receive reward $X_{a_t,t} \sim \nu_{a_t}$

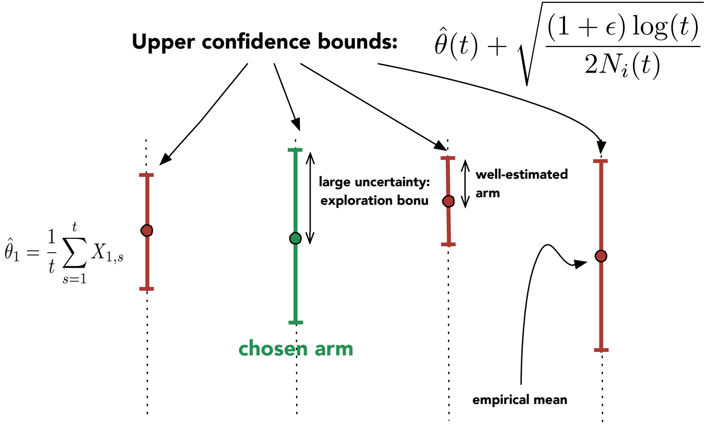
Goal: maximize $\mathbb{E}[\sum_{t=1}^{T} X_{a_t,t}]$

Dilemma: exploration vs *exploitation*



A successful approach: UCB algorithm (Auer et al., 2002)

- *Initialization:* pull each arm once
- Then:



The Profitable Bandit problem

At each time $t = 1, ..., T$

- Choose arms $A_t \subset \{1, ..., K\}$
- Observe rewards $X_{a,c,t} \sim \nu_a$ for all $a \in A_t$, $c \in \{1, \ldots, C_a(t)\}$

Objective

maximize $S_T := \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in A_t} \sum_{c=1}^{C_a(t)} (X_{a,c,t} - \tau_a)]$

The Profitable Bandit problem

At each time $t = 1, ..., T$

- Choose arms $A_t \subset \{1, ..., K\}$
- Observe rewards $X_{a,c,t} \sim \nu_a$ for all $a \in A_t$, $c \in \{1, \ldots, C_a(t)\}$

Objective

maximize $S_T := \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in A_t} \sum_{c=1}^{C_a(t)} (X_{a,c,t} - \tau_a)]$

Hence, optimal choice: $A^* = \{a \in \{1, \ldots, K\}, \Delta_a > 0\}$
with $\Delta_a = \mu_a - \tau_a$ and $\mu_a = \mathbb{E}[X_{a,1,1}]$.

The Profitable Bandit problem

At each time $t = 1, ..., T$

- Choose arms $A_t \subset \{1, ..., K\}$
- Observe rewards $X_{a,c,t} \sim \nu_a$ for all $a \in A_t$, $c \in \{1, \ldots, C_a(t)\}$

Objective

maximize $S_T := \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in A_t} \sum_{c=1}^{C_a(t)} (X_{a,c,t} - \tau_a)]$

Hence, optimal choice: $A^* = \{a \in \{1, \ldots, K\}, \Delta_a > 0\}$
with $\Delta_a = \mu_a - \tau_a$ and $\mu_a = \mathbb{E}[X_{a,1,1}]$.
Equivalently, minimize the expected regret

$$R_T = \sum_{a \in A^*} \Delta_a \tilde{C}_a(T) - S_T$$
$$= \sum_{a \in A^*} \Delta_a \left(\tilde{C}_a(T) - \mathbb{E}[N_a(T)]\right) + \sum_{a \notin A^*} |\Delta_a| \mathbb{E}[N_a(T)],$$

where $\tilde{C}_a(T) = \mathbb{E}[\sum_{t=1}^{T} C_a(t)]$, $N_a(T) = \sum_{t=1}^{T} C_a(t) \mathbb{I}\{a \in A_t\}$.

Outline

Lower bound

Theorem

If the ν_a's belong to an one-dimensional exponential family, for all uniformly efficient strategies, for all non-profitable arms a such that $\mu_a < \tau_a$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \tau_a)},$$

with d the KL-divergence of the family parametrized by the mean: $d(\mu_a, \mu_{a'}) = KL(\nu_a, \nu_{a'})$.

Lower bound

Theorem

If the ν_a's belong to an one-dimensional exponential family, for all uniformly efficient strategies, for all non-profitable arms a such that $\mu_a < \tau_a$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \tau_a)},$$

with d the KL-divergence of the family parametrized by the mean: $d(\mu_a, \mu_{a'}) = KL(\nu_a, \nu_{a'})$.

Consequence:

$$R_T \gtrsim \sum_{a \notin A^*} \frac{|\Delta_a|}{d(\mu_a, \tau_a)} \log T.$$

Outline

Index policy

An index policy is fully characterized by the choice of index $u_a(t)$.

Algorithm 1 Generic index policy

Require: time horizon T, thresholds $(\tau_a)_{a \in \{1,\ldots,K\}}$

1: Pull all arms: $A_1 = \{1, \ldots, K\}$

2: **for** $t = 1$ **to** $T - 1$ **do**

3: Compute $u_a(t)$ for all arms $a \in \{1, \ldots, K\}$

4: Choose $A_{t+1} \leftarrow \{a \in \{1, \ldots, K\}, u_a(t) \geq \tau_a\}$

5: **end for**

Three index policies

- kl-UCB-4P

$$u_a(t) = \sup\left\{ q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t \right\}$$

Three index policies

- kl-UCB-4P

$$u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t\right\}$$

- Bayes-UCB-4P

$$u_a(t) = Q(1 - 1/(t(\log t)^c); \lambda_a^{t-1}),$$

with λ_a^{t-1} the posterior distribution on μ_a after round $t-1$.

Three index policies

- kl-UCB-4P

$$u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t\right\}$$

- Bayes-UCB-4P

$$u_a(t) = Q(1 - 1/(t(\log t)^c); \lambda_a^{t-1}),$$

  with λ_a^{t-1} the posterior distribution on μ_a after round $t-1$.

- Thompson-Sampling-4P

$$u_a(t) = \mu(\theta_{a,t}),$$

  where $\theta_{a,t} \sim \pi_a^{t-1}$ with π_a^{t-1} the posterior distribution on θ_a after round $t-1$.

Outline

Upper bound

Theorem

For the three policies defined above (kl-UCB-4P, Bayes-UCB-4P, TS-4P),

$$R_T \leq \sum_{a \notin A^*} \frac{c_a^+}{c_a^-} \frac{|\Delta_a|}{d(\mu_a, \tau_a)} \log T + o(\log \log T),$$

where for all $t \geq 1$: $c_a^- \leq C_a(t) \leq c_a^+$.

Upper bound

Theorem

For the three policies defined above (kl-UCB-4P, Bayes-UCB-4P, TS-4P),

$$R_T \leq \sum_{a \notin A^*} \frac{c_a^+}{c_a^-} \frac{|\Delta_a|}{d(\mu_a, \tau_a)} \log T + o(\log \log T),$$

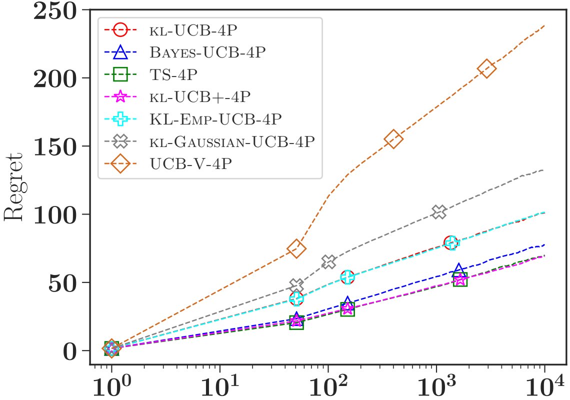
where for all $t \geq 1$*:* $c_a^- \leq C_a(t) \leq c_a^+$.

Conclusion: the three algorithms are asymptotically optimal when $C_a(1) = \cdots = C_a(T)$ for all $a \notin A^*$.

Outline

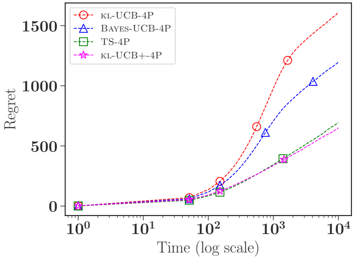
Bernoulli distributions

- $T = 10^4$
- $K = 5$ arms, $\nu_a = \text{Bernoulli}(\mu_a)$
- (μ_a, τ_a): $(0.1, 0.2)$, $(0.3, 0.2)$, $(0.5, 0.4)$, $(0.5, 0.6)$, $(0.7, 0.8)$
- $C_a(t) - 1 \sim \text{Poisson}(a + 1)$

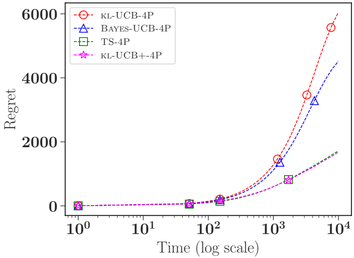


Poisson and Exponential distributions

Poisson



Exponential



Thank you!