Profitable Bandits

Mastane Achab[1] Stephan Clémençon[1] Aurélien Garivier[2]

[1] LTCI, Télécom ParisTech, Paris, France
[2] UMPA & LIP, ENS de Lyon, Lyon, France

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 $a \in A_t$, $c \in \{1, \dots, C_a(t)\}$

$*$ Objective: maximize $S_T = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in A_t} \sum_{c=1}^{C_a(t)} (X_{a,c,t} - \tau_a)\right]$.
$*$ Optimal choice: $A^* = \{a \in \{1, \dots, K\}, \Delta_a > 0\}$ 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 $\tilde{C}_a(T) = \mathbb{E}[\sum_{t=1}^{T} C_a(t)]$ and $N_a(T) = \sum_{t=1}^{T} C_a(t)\mathbb{I}\{a \in A_t\}$.

Lower bound

Theorem 1. *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:

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

Index policies

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

Generic index policy

Require: time horizon T, thresholds $(\tau_a)_{a \in \{1, \dots, K\}}$.
Pull all arms: $A_1 = \{1, \dots, K\}$.
for $t = 1$ **to** $T - 1$ **do**
  Compute $u_a(t)$ for all arms $a \in \{1, \dots, K\}$.
  Choose $A_{t+1} \leftarrow \{a \in \{1, \dots, K\}, u_a(t) \geq \tau_a\}$.
end for

We consider the three following 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 post. distrib. 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 post. distrib. on θ_a after round $t-1$.

Upper bound

Theorem 2. *For kl-UCB-4P, Bayes-UCB-4P and TS-4P:*

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

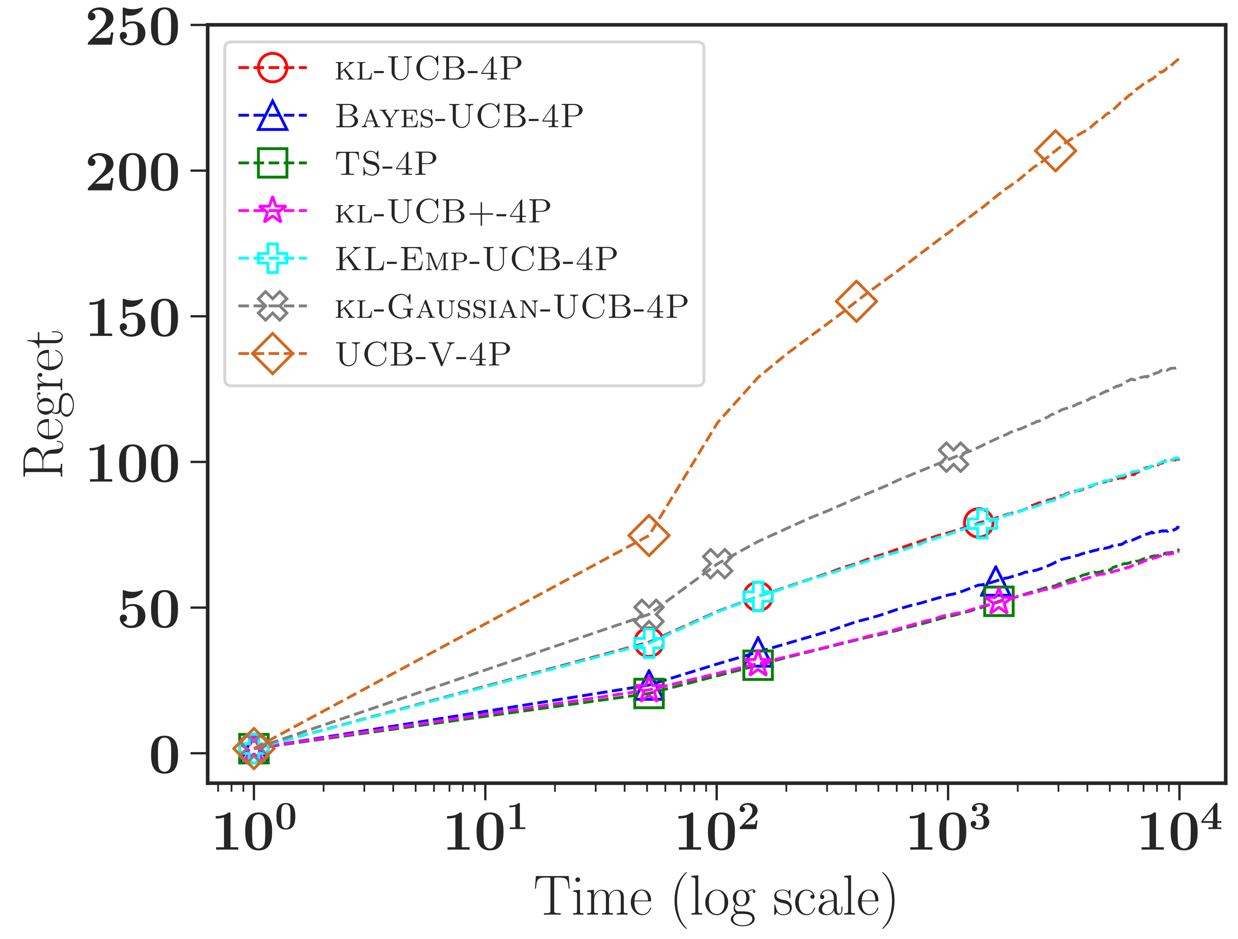
where c_a^- and c_a^+ are constants such that for all $t \geq 1$: $c_a^- \leq C_a(t) \leq c_a^+$.

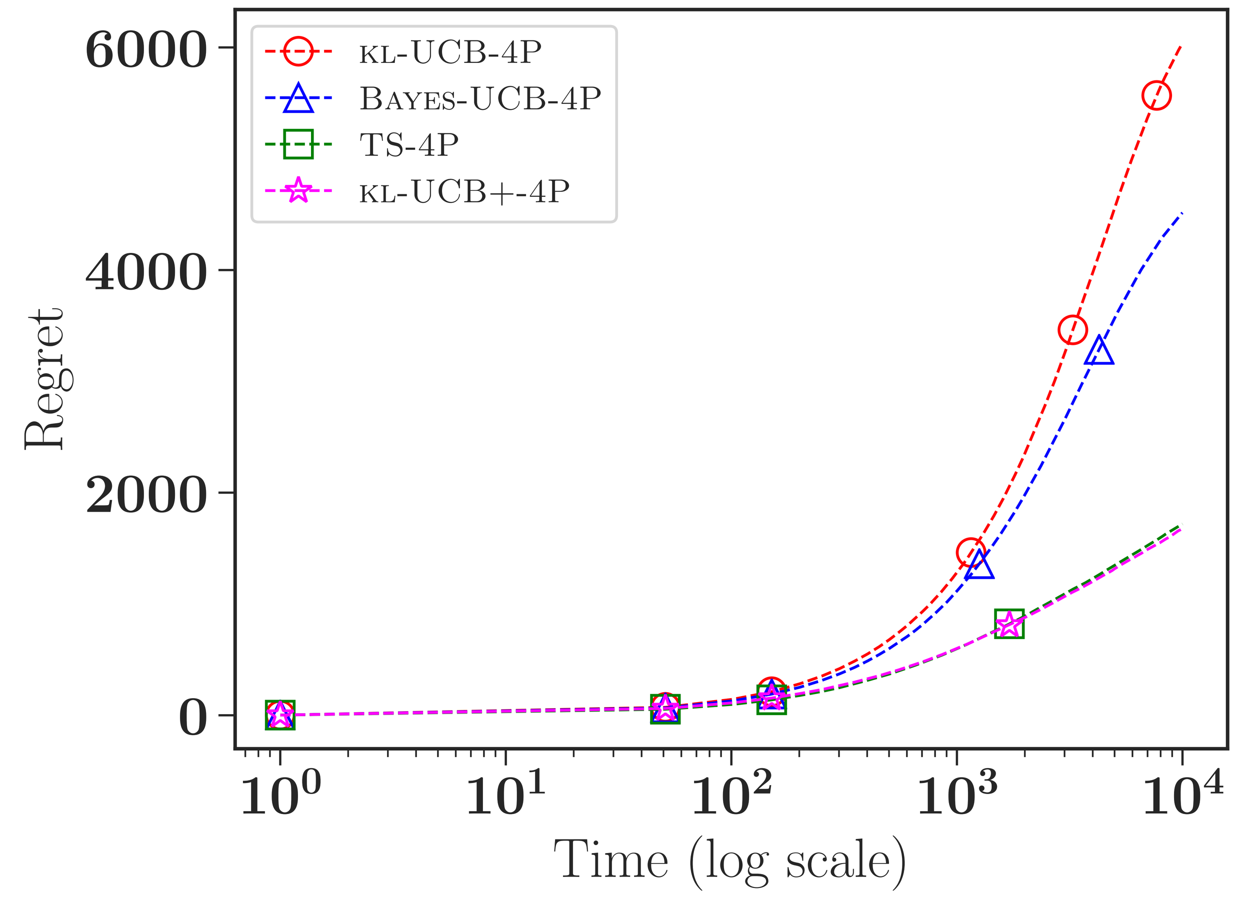
$*$ Conclusion: the three algorithms are asymptotically optimal when $C_a(1) = \dots = C_a(T)$ for all $a \notin A^*$.

Numerical experiments

$*$ Bernoulli scenario (left figure). Problem parameters: $T = 10^4$, $K = 5$, $C_a(t) - 1 \sim \text{Poisson}(a+1)$, (μ_a, τ_a): $(0.1, 0.2)$, $(0.3, 0.2)$, $(0.5, 0.4)$, $(0.5, 0.6)$, $(0.7, 0.8)$.

$*$ Exponential scenario (right figure). Same parameters except for the μ_a's and τ_a's. (μ_a, τ_a): $(1, 1.1)$, $(2, 1.9)$, $(3, 3.1)$, $(4, 3.9)$, $(5, 5.1)$.





$*$ Interpretation: the best performing policies are those adapting to the parametric family of the reward distributions: through the Kullback-Leibler divergence for kl-UCB-4P or through prior distributions for Bayes-UCB-4P and TS-4P.

References

- Garivier, A. and Cappé, O. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. Conference On Learning Theory, 2011.
- Kaufmann, E. On Bayesian index policies for sequential resource allocation. Annals of Statistics, 2017.
- Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933.
- Korda, N. and Kaufmann, E. and Munos, R. Thompson sampling for 1-dimensional exponential family bandits. Advances in Neural Information Processing Systems, 2013.