Max K-armed Bandits: on the ExtremeHunter Algorithm and an Alternative Approach

Mastane Achab[1] Stéphan Clémençon[1] Aurélien Garivier[2] Anne Sabourin[1] Claire Vernade[1]

[1] LTCI, Télécom ParisTech, Université Paris-Saclay [2] IMT, Université de Toulouse

Max K-Armed Bandits

- The max K-armed bandit problem (Cicirello and Smith, 2005) is a sequential decision-making problem in an uncertain environment. At each time $t = 1, ..., n$
  - choose arm $k_t \in \{1, ..., K\}$
  - observe reward $X_{k_t,t} \sim \nu_{k_t}$.
- Objective: maximize $\mathbb{E}\left[\max_{1\le t\le n} X_{k_t,t}\right]$.
- Or equivalently: minimize the *expected extreme regret*

$$\mathbb{E}\left[R_n\right] \triangleq \mathbb{E}\left[\max_{1\le t\le n} X_{k^*,t}\right] - \mathbb{E}\left[\max_{1\le t\le n} X_{k_t,t}\right] ,$$

where $k^* \triangleq \arg\max_{1\le k\le K} \mathbb{E}[\max_{1\le t\le n} X_{k,t}]$ is the *optimal arm.*

Second-order Pareto

- An (α, β, C, C')-second order Pareto with cdf F verifies $\forall x \ge 0$

$$|1 - Cx^{-\alpha} - F(x)| \le C'x^{-\alpha(1+\beta)} .$$

- As in [1], rewards $X_{k,t} \sim \nu_k$ with ν_k an $(\alpha_k, \beta_k, C_k, C')$-second order Pareto distribution, $\alpha_k > 1$, $\beta_k > 0$, $C_k > 0$ and $C' > 0$.

Theorem 1. *If $\alpha > 1$ then*

$$\left|\mathbb{E}\left[\max_{1\le t\le n} X_t\right] - (nC)^{1/\alpha}\Gamma\left(1 - 1/\alpha\right)\right| = \mathcal{O}\left(n^{-(\min(1,\beta)-1/\alpha)}\right) ,$$

sharper than $\mathcal{O}\left(n^{\frac{1}{(1+\beta)\alpha}}\right)$ *in [1].*

ExtremeETC Algorithm

We propose ExtremeETC, an *Explore-Then-Commit* version of ExtremeHunter [1]. Both use the *optimism-in-the-face-of-uncertainty* principle through indices

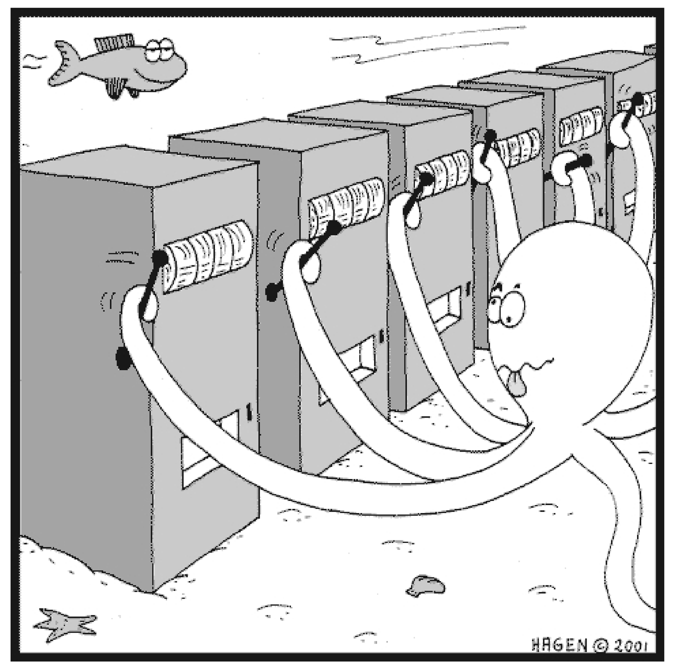
$$B_k \triangleq \left(n(\widehat{C_k} + \Lambda_2)\right)^{\widehat{1/\alpha_k}+\Lambda_1}\Gamma(1 - \widehat{1/\alpha_k} - \Lambda_1) \quad (1)$$
$$\left(\ge \mathbb{E}\left[\max_{1\le t\le T} X_{k,t}\right] \text{ with high probability}\right) .$$

ExtremeETC vs ExtremeHunter

1: **Input:** n: time horizon, K: number of arms, $b > 0$ such that $b \le \min_k \beta_k$.
2: **Initialize:** Pull N times each arm k and compute index B_k (see Eq. (1)).

ExtremeETC	ExtremeHunter
3: $k_0 = \arg\max_k B_k$	3: **for** $t > KN$ **do**
4: **for** $t > KN$ **do**	4: Pull $k_t = \arg\max_k B_k$.
5: Pull arm k_0.	5: Update index B_{k_t}.
6: **end for**	6: **end for**

Complexity	Ex.ETC	Ex.Hunt.
Time	$\mathcal{O}\left(K(\log n)^{\frac{2(2b+1)}{b}}\right)$	$\mathcal{O}(n^2)$
Memory	$\mathcal{O}\left(K(\log n)^{\frac{2(2b+1)}{b}}\right)$	$\mathcal{O}(n)$



Tight regret bounds

Theorem 2. *(i) Upper bound for* ExtremeETC *and* ExtremeHunter

$$\mathbb{E}[R_n] = \mathcal{O}\Big((\log n)^{\frac{2(2b+1)}{b}} n^{-(1-1/\alpha_{k^*})} + n^{-(b-1/\alpha_{k^*})}\Big) ,$$

sharper than $\mathcal{O}\left(n^{\frac{1}{(1+b)\alpha_{k^*}}}\right)$ *in [1].*

(ii) Lower bound for any algorithm pulling each arm at least N times

$$\mathbb{E}[R_n] = \Omega\left((\log n)^{\frac{2(2b+1)}{b}} n^{-(1-1/\alpha_{k^*})}\right) .$$

When $b \ge 1$, (i) and (ii) are tight !

Reduction to Multi-Armed Bandits (MAB)

- Truncated rewards: $Y_{k,t} \triangleq X_{k,t}\mathbb{1}_{\{X_{k,t}>u\}}$.
- $\mathbb{E}[Y_{k,1}] \sim_{u\to\infty} C_k\left(1 + \frac{1}{\alpha_k - 1}\right)u^{-\alpha_k+1}$.
- For $u > \max\left(1, \left(\frac{2C'}{C_{(1)}}\right)^{\frac{1}{b}}, \left(\frac{3C_{(K)}}{C_{(1)}}\right)^{\frac{1}{\alpha_{(2)}-\alpha_{(1)}}}\right)$ and n large enough

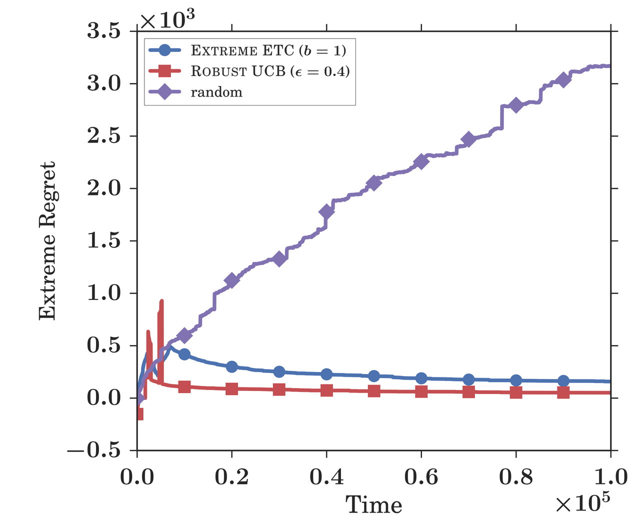
$$\arg\max_{1\le k\le K} \mathbb{E}[Y_{k,1}] = \arg\min_{1\le k\le K} \alpha_k = k^* .$$

- MAB objective: maximize $\mathbb{E}\left[\sum_{t=1}^n Y_{k_t,t}\right]$.
- We use Robust UCB with truncated mean estimator [4]
  - parameters: $\epsilon < \min_{1\le k\le K} \alpha_k - 1$, $v \ge \max_{k\in[K]} \mathbb{E}\left[|Y_{k,1}|^{1+\epsilon}\right]$
  - $\mathbb{E}\left[\# \text{ pulls arm } k \ne k^*\right] = \mathcal{O}(\log n) < N$.

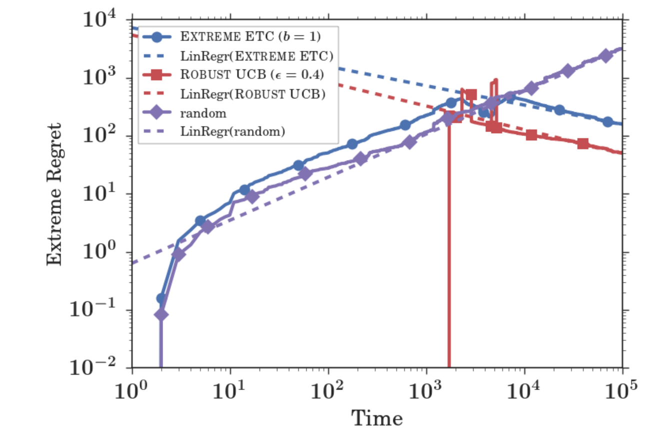
Experiments

- time horizon $n = 10^5$
- $K = 3$ exact Pareto distributions ($\beta = +\infty$)

Arm	1	$k^* = 2$	3
α	15	1.5	10
C	10^8	1	10^5
$\mathbb{E}[X]$	3.7	3	3.5
$\mathbb{E}[\max_{1\le t\le n} X_t]$	7.7	$5.8 \cdot 10^3$	11



(a) ExtremeETC vs Robust UCB vs uniform random



(b) same in log-log plot

Figure 1: Extreme regret averaged over 1000 independent trajectories.

- Fig. 1b: linear regression for ExtremeETC over $t = 5 \cdot 10^4, \ldots, 10^5$ has slope ≈ -0.333 **→ validation of Theorem 2 !**

Estimation of $1/\alpha$ and C (see resp. [2] and [3])

Assume $T \ge N \triangleq A(\log n)^{\frac{2(2b+1)}{b}}$ with b known s.t. $b \le \beta$.

- $\widehat{1/\alpha} \triangleq \min\left(1, \left[\log\left(\frac{\sum_{t=1}^T \mathbb{1}_{\{X_t > e^r\}}}{\sum_{t=1}^T \mathbb{1}_{\{X_t > e^{r+1}\}}}\right)\right]^{-1}\right)$
- $\widehat{C} \triangleq T^{-\frac{2b}{2b+1}} \sum_{t=1}^T \mathbb{1}_{\{X_t \ge T^{\widehat{1/\alpha}/(2b+1)}\}}$

With probability larger than $1 - \delta$,

- $\left|\widehat{1/\alpha} - 1/\alpha\right| \le \Lambda_1(T) \triangleq D\sqrt{\log(1/\delta)}T^{-\frac{b}{2b+1}}$
- $\left|\widehat{C} - C\right| \le \Lambda_2(T) \triangleq E\sqrt{\log(T/\delta)}\log(T)T^{-\frac{b}{2b+1}}$.

References

References

[1] Alexandra Carpentier and Michal Valko. Extreme bandits (2014).

[2] Alexandra Carpentier and Arlene KH Kim. Adaptive and minimax optimal estimation of the tail coefficient (2014).

[3] Alexandra Carpentier, Arlene KH Kim, et al. Honest and adaptive confidence interval for the tail coefficient in the pareto model (2014).

[4] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail (2013).