Efficient Algorithms for Extreme Bandits

Dorian Baudry[1] Yoan Russac[2], Emilie Kaufmann[1]
[1] INRIA & Université de Lille [2] CNRS & ENS

Problem and setting

- Learner interacts with K unknown arms denoted $\nu_1, \dots, \nu_K$.
- $X_{k,t}$ obtained by pulling arm k at time t.
- The learner seeks to collect the **largest possible reward**.
- **Minimizing the extreme regret** which for a policy π that selects arm I_t at time t is defined by:

$$\mathcal{R}_T^\pi = \max_{k \le K} \mathbb{E}\left[\max_{t \le T} X_{k,t}\right] - \mathbb{E}_\pi\left[\max_{t \le T} X_{I_t,t}\right] .$$

Two notions of convergence:

- **Weakly Vanishing Regret**: $\mathcal{R}_T^\pi = \underset{T\to\infty}{o}\left(\max_{k\le K}\mathbb{E}[\max_{t\le T} X_{k,t}]\right)$.
- **Strongly Vanishing Regret**: $\lim_{T\to\infty} \mathcal{R}_T^\pi = 0$.

Challenge

- **Relaxing parametric assumption on the distributions while obtaining strong theoretical guarantees**
  1. Some works assume that the distribution are known (Frechet, Gumbel)
  2. Other works have a semi-parametric assumption (second-order Pareto)
  3. Some works with weaker assumptions but hard to obtain guarantees.
- **Reducing computational and storage cost compared to existing approaches.**

Algorithm	Memory	Time
Extreme Hunter	T	$\mathcal{O}(T^2)$
MaxMedian	T	$\mathcal{O}(KT\log T)$
QoMax-SDA	$\mathcal{O}((\log T)^2)$	$\mathcal{O}(KT\log T)$
Extreme ETC	$\mathcal{O}\left(K(\log T)^3\right)$	$\mathcal{O}\left(K(\log T)^6\right)$
QoMax-ETC	$\mathcal{O}(K(\log T)^2)$	$\mathcal{O}(K(\log T)^3)$

Table: Average time and storage complexities of Extreme Bandit algorithms for a time horizon T.

Dominating Tail

Definition 1 (Exponential or polynomial tails). *Let ν be a distribution of survival function G. **(1)** If there exists $C > 0$ and $\lambda > 1$ such that $G(x) \sim Cx^{-\lambda}$ we say that ν has a **polynomial tail**. **(2)** If there exists $C > 0, \lambda \in \mathbb{R}^+$ such that $G(x) \sim C\exp(-\lambda x)$ we say that ν has an **exponential tail**.*

Definition 2 (Dominating tail). *Let G_1 and G_2 be the survival functions of two distributions ν_1 and ν_2. We say that the tail of ν_1 **dominates** the tail of ν_2 (we write $\nu_1 \succ \nu_2$) if there exists $C > 1$ and $x \in \mathbb{R}$ such that for all $y > x$, $G_1(y) > CG_2(y)$.*

Quantile of Maxima (QoMax) estimator

→ Inspired by **Median of Means estimator**.

→ Learner separates the data into **batches** of **equal sizes** and compute the quantile of order q of the maxima over the different batches. With $N = b \times n$ data points, the learner allocates the data in b batches of size n and:

1. find the maximum of each batch
2. compute the quantile q over the b maxima.

→ $\bar{X}^q_{k,n,b}$ is the QoMax of order q computed from b batches of size n of i.i.d. replications from arm k.

QoMax-ETC

For $k \le K$:

Pull arm k, $b_T \times n_T$ times.

Allocate the data in b_T batches of size n_T. Compute their QoMax, $\bar{X}^q_{k,n_T,b_T}$

For $t = K \times n_T \times b_T + 1, \dots, T$: Pull arm $I_T = \text{argmax}_k \bar{X}^q_{k,n_T,b_T}$

QoMaX-SDA

A **round-based** algorithm based on three ingredients

1. Selection of a leader: arm that has been pulled the most: $\ell(r) = \text{argmax}_{k\le K}\, n_k(r)$.
2. Duels between the leader and the $K-1$ remaining arms: comparison of the QoMax of the challenger using its entire history and the **QoMax of the leader on a subsample of its history**.
3. Data collection procedure.

Beginning of round r:

- $\mathcal{X}_k^r$ the history of arm k with $b_k(r)$ batches of size $n_k(r)$. $f(r)$ represents the *sampling obligation* at round r.
- **Duel**: Arm $k \in \mathcal{A}_{r+1}$ (pulled arms) if (1) it wins its duel OR (2) undersampled i.e. $n_k(r) \le f(r)$.
- We assume that $b_k(r)$ depends only on its number of queries $n_k(r)$ i.e. $b_k(r) = B(n_k(r))$ for some function B.

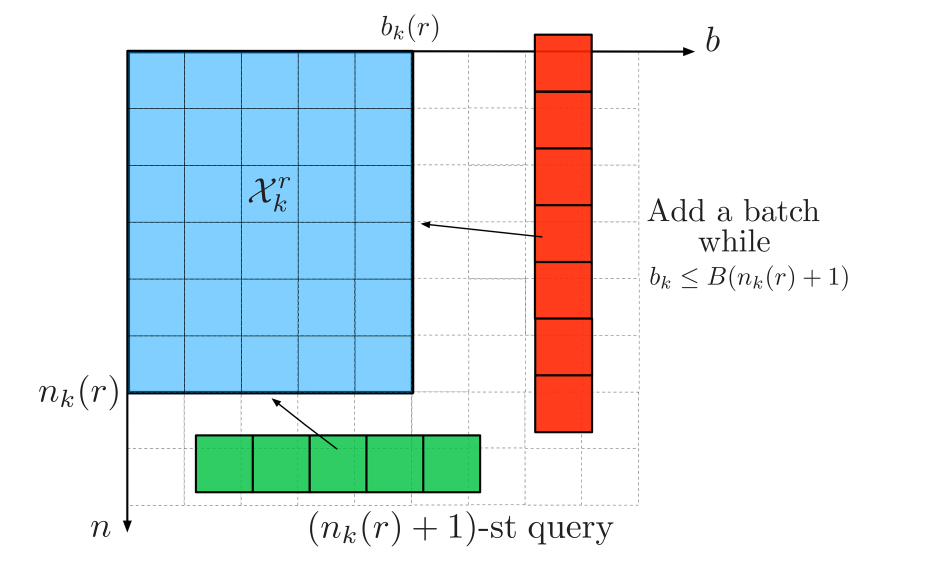

Figure 1: Illustration of the CollectData procedure at round r for a challenger $k \in \mathcal{A}_{r+1}$.

Concentration QoMax

Theorem 1 (Comparison of QoMax). *Let ν_1 and ν_2 be two distributions satisfying $\nu_1 \succ \nu_2$ and $q \in (0,1)$. Then, **there exists** a sequence (x_n), a constant $c > 0$, and an integer $n_{\nu_1,\nu_2,q}$ such that for $n \ge n_{\nu_1,\nu_2,q}$,*

$$\max\left\{\mathbb{P}(\bar{X}^q_{1,n,b} \le x_n), \mathbb{P}(\bar{X}^q_{2,n,b} \ge x_n)\right\} \le \exp(-cb) .$$

*If the tails are furthermore either polynomial or exponential with a **positive tail gap**, then the result holds **for any** $c > 0$ and n larger than some $n_{c,\nu_1,\nu_2,q}$.*

→ $\mathbb{P}(\bar{X}^q_{1,n,b} \le \bar{X}^q_{2,n,b}) \le 2\exp(-cb)$: the comparison of QoMax computed with the same parameters will not be in favor of the dominating arm with a probability that **decreases exponentially with the batch size**

Theoretical Guarantees

Theorem 2 (Vanishing regret of QoMax-ETC). *Consider a bandit with $\nu_1 \succ \nu_k$ for $k \ne 1$. Under proper assumption, for any quantile $q \in (0,1)$ and any sequence (b_T, n_T) satisfying $\frac{b_T}{\log(T)} \to +\infty$ and $n_T \to +\infty$, the regret of QoMax-ETC with parameters (q, b_T, n_T) is **vanishing in the strong sense**.*

Theoretical Guarantees

Theorem 3 (Upper bound on the regret of QoMax-SDA). *For any quantile q, any $\gamma > 0$, defining the parameters of QoMax-SDA as $B(n) = n^\gamma$ and $f(r) = (\log r)^{\frac{1}{\gamma}}$.*

The regret of QoMax-SDA is:

1. *Vanishing in the **strong sense** for exponential tails.*
2. *Vanishing in the **weak sense** for polynomial tails.*

Numerical Results

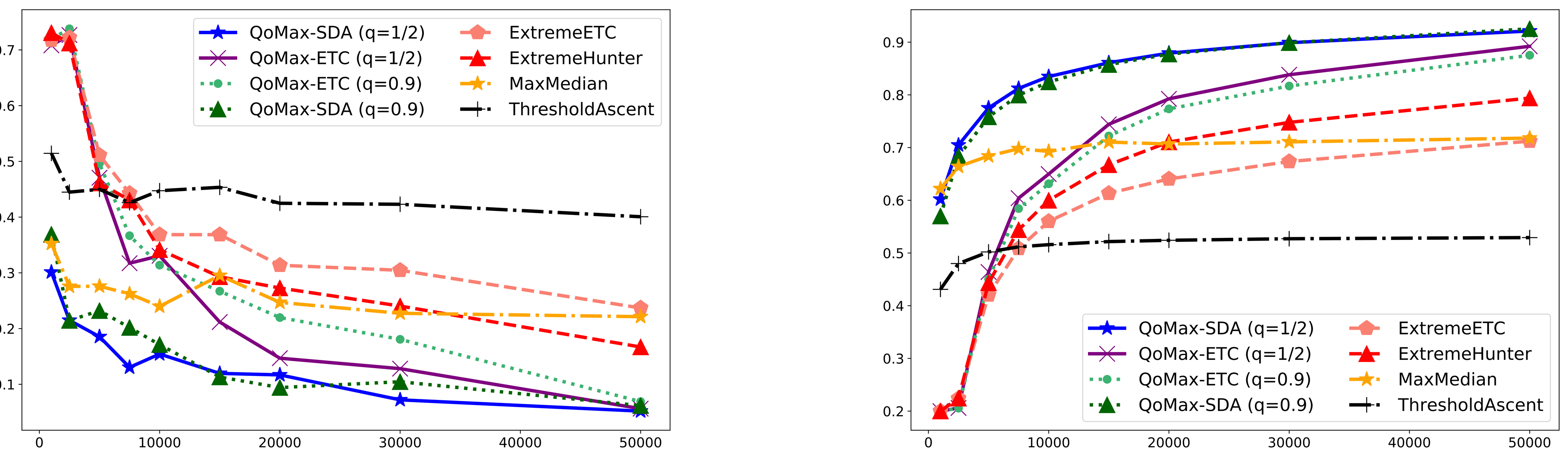

Figure 2: Proxy Empirical Regret **(I)** and Percentage of best arm pulls **(II)** averaged over 10^4 independent trajectories for $T \in \{10^3, 2.5\times10^3, 5\times10^3, 7.5\times10^3, 9\times10, 10^4, 1.5\times10^4, 2\times10^4, 3\times10^4, 5\times10^4\}$.