Max K-armed bandit:
On the ExtremeHunter algorithm and beyond

Mastane Achab[1], Stephan Clémençon[1], Aurélien Garivier[2], Anne Sabourin[1],
and Claire Vernade[1]

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

Abstract. This paper is devoted to the study of the *max K-armed bandit problem*, which consists in sequentially allocating resources in order to detect extreme values. Our contribution is twofold. We first significantly refine the analysis of the ExtremeHunter algorithm carried out in Carpentier and Valko (2014), and next propose an alternative approach, showing that, remarkably, Extreme Bandits can be reduced to a classical version of the bandit problem to a certain extent. Beyond the formal analysis, these two approaches are compared through numerical experiments.

1 Introduction

In a classical multi-armed bandit (MAB in abbreviated form) problem, the objective is to find a strategy/policy in order to sequentially explore and exploit K sources of gain, referred to as *arms*, so as to maximize the expected cumulative gain. Each arm $k \in \{1, \ldots, K\}$ is characterized by an unknown probability distribution ν_k. At each round $t \geq 1$, a strategy π picks an arm $I_t = \pi((I_1, X_{I_1,1}), \ldots, (I_{t-1}, X_{I_{t-1},t-1}))$ and receives a random reward $X_{I_t,t}$ sampled from distribution ν_{I_t}. Whereas usual strategies aim at finding and exploiting the arm with highest expectation, the quantity of interest in many applications such as medicine, insurance or finance may not be the sum of the rewards, but rather the *extreme* observations (even if it might mean replacing loss minimization by gain maximization in the formulation of the practical problem). In such situations, classical bandit algorithms can be significantly sub-optimal: the "best" arm should not be defined as that with highest expectation, but as that producing the maximal values. This setting, referred to as *extreme bandits* in Carpentier and Valko (2014), was originally introduced by Cicirello and Smith (2005) by the name of *max K-armed bandit problem*. In this framework, the goal pursued is to obtain the highest possible reward during the first $n \geq 1$ steps. For a given arm k, we denote by

$$G_n^{(k)} = \max_{1 \leq t \leq n} X_{k,t}$$

the maximal value taken until round $n \geq 1$ and assume that, in expectation, there is a unique optimal arm

$$k^* = \arg\max_{1 \leq k \leq K} \mathbb{E}[G_n^{(k)}] \; .$$

The expected *regret* of a strategy π is here defined as

$$\mathbb{E}[R_n] = \mathbb{E}[G_n^{(k^*)}] - \mathbb{E}[G_n^{(\pi)}], \tag{1}$$

where $G_n^{(\pi)} = \max_{1 \leq t \leq n} X_{I_t,t}$ is the maximal value observed when implementing strategy π. When the supports of the reward distributions (*i.e.* the ν_k's) are bounded, no-regret is expected provided that every arm can be sufficiently explored, refer to Nishihara et al. (2016) (see also David and Shimkin (2016) for a PAC approach). If infinitely many arms are possibly involved in the learning strategy, the challenge is then to explore and exploit optimally the unknown reservoir of arms, see Carpentier and Valko (2015). When the rewards are unbounded in contrast, the situation is quite different: the best arm is that for which the maximum $G_n^{(k)}$ tends to infinity faster than the others. In Nishihara et al. (2016), it is shown that, for unbounded distributions, no policy can achieve no-regret without restrictive assumptions on the distributions. In accordance with the literature, we focus on a classical framework in extreme value analysis. Namely, we assume that the reward distributions are *heavy-tailed*. Such Pareto-like laws are widely used to model extremes in many applications, where a conservative approach to risk assessment might be relevant (*e.g.* finance, environmental risks). Like in Carpentier and Valko (2014), rewards are assumed to be distributed as second order Pareto laws in the present article. For the sake of completeness, we recall that a probability law with cdf $F(x)$ belongs to the (α, β, C, C')-second order Pareto family if, for every $x \geq 0$,

$$|1 - Cx^{-\alpha} - F(x)| \leq C'x^{-\alpha(1+\beta)} \, , \tag{2}$$

where α, β, C and C' are strictly positive constants, see *e.g.* Resnick (2007). In this context, Carpentier and Valko (2014) have proposed the EXTREMEHUNTER algorithm to solve the *extreme bandit* problem and provided a regret analysis.

The contribution of this paper is twofold. First, the regret analysis of the EXTREMEHUNTER algorithm is significantly improved, in a nearly optimal fashion. This essentially relies on a new technical result of independent interest (see Theorem 1 below), which provides a bound for the difference between the expectation of the maximum among independent realizations X_1, ..., X_T of a (α, β, C, C')-second order Pareto distribution, $\mathbb{E}[\max_{1 \leq i \leq T} X_i]$ namely, and its rough approximation $(TC)^{1/\alpha}\Gamma(1 - 1/\alpha)$. As a by-product, we propose a more simple EXPLORE-THEN-COMMIT strategy that offers the same theoretical guarantees as EXTREMEHUNTER. Second, we explain how extreme bandit can be reduced to a classical bandit problem to a certain extent. We show that a regret-minimizing strategy such as ROBUST-UCB (see Bubeck et al. (2013)), applied on correctly left-censored rewards, may also reach a very good performance. This

claim is supported by theoretical guarantees on the number of pulls of the best arm k^* and by numerical experiments both at the same time. From a practical angle, the main drawback of this alternative approach consists in the fact that its implementation requires some knowledge of the complexity of the problem (*i.e.* of the gap between the first-order Pareto coefficients of the first and second arms). In regard to its theoretical analysis, efficiency is proved for large horizons only.

This paper is organized as follows. Section 2 presents the technical result mentioned above, which next permits to carry out a refined regret analysis of the EXTREMEHUNTER algorithm in Section 3. In Section 4, the regret bound thus obtained is proved to be nearly optimal: precisely, we establish a lower bound under the assumption that the distributions are close enough to Pareto distributions showing the regret bound is sharp in this situation. In Section 5, reduction of the extreme bandit problem to a classical bandit problem is explained at length, and an algorithm resulting from this original view is then described. Finally, we provide a preliminary numerical study that permits to compare the two approaches from an experimental perspective. Due to space limitations, certain technical proofs are deferred to the Supplementary Material[3].

2 Second-order Pareto distributions: approximation of the expected maximum among i.i.d. realizations

In the extreme bandit problem, the key to controlling the behavior of explore-exploit strategies is to approximate the expected payoff of a fixed arm $k \in \{1, \ldots, K\}$. The main result of this section, stated in Theorem 1, provides such control: it significantly improves upon the result originaly obtained by Carpentier and Valko (2014) (see Theorem 1 therein). As shall be next shown in Section 3, this refinement has substantial consequences on the regret bound.

In Carpentier and Valko (2014), the distance between the expected maximum of independent realizations of a (α, β, C, C')-second order Pareto and the corresponding expectation of a Fréchet distribution $(TC)^{1/\alpha}\Gamma(1 - 1/\alpha)$ is controlled as follows:

$$\left| \mathbb{E}\left[\max_{1 \leq i \leq T} X_i \right] - (TC)^{1/\alpha}\Gamma(1 - 1/\alpha) \right| \leq \frac{4D_2 C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T^{\beta-1/\alpha}} + (2C'T)^{\frac{1}{(1+\beta)\alpha}} .$$

Notice that the leading term of this bound is $(2C'T)^{1/((1+\beta)\alpha)}$ as $T \to +\infty$. Below, we state a sharper result where, remarkably, this (exploding) term disappears, the contribution of the related component in the approximation error decomposition being proved as (asymptotically) negligible in contrast.

Theorem 1. (FRÉCHET APPROXIMATION BOUND) *If* X_1, $\ldots$, X_T *are i.i.d. r.v.'s drawn from a* (α, β, C, C')-*second order Pareto distribution with* $\alpha > 1$

---

[3] See the full-length paper: http://arxiv.org/abs/1707.08820.

and $T \geq Q_1$, where Q_1 is the constant depending only on α, β, C and C' given in Eq. (3) below, then,

$$\left| \mathbb{E}\left[\max_{1 \leq i \leq T} X_i \right] - (TC)^{1/\alpha} \Gamma(1 - 1/\alpha) \right|$$
$$\leq \frac{4D_2 C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C' D_{\beta+1}}{C^{\beta+1-1/\alpha} T^{\beta-1/\alpha}} + 2(2C'T)^{\frac{1}{(1+\beta)\alpha}} e^{-HT^{\frac{\beta}{\beta+1}}}$$
$$= \underset{T \to \infty}{o}\left(T^{1/\alpha}\right),$$

where $H = C(2C')^{1/(\alpha(1+\beta))}/2$. In particular, if $\beta \geq 1$, we have:

$$\left| \mathbb{E}\left[\max_{1 \leq i \leq T} X_i \right] - (TC)^{1/\alpha} \Gamma(1 - 1/\alpha) \right| = o(1) \text{ as } T \to +\infty.$$

We emphasize that the bound above shows that the distance of $\mathbb{E}[\max_{1 \leq i \leq T} X_i]$ to the Fréchet mean $(TC)^{1/\alpha} \Gamma(1 - \frac{1}{\alpha})$ actually vanishes as $T \to \infty$ as soon as $\beta \geq 1$, a property that shall be useful in Section 3 to study the behavior of learning algorithms in the extreme bandit setting.

Proof. Assume that $T \geq Q_1$, where

$$Q_1 = \frac{1}{2C'} \max \left\{ (2C'/C)^{(1+\beta)/\beta}, \ (8C)^{1+\beta} \right\}. \tag{3}$$

As in the proof of Theorem 1 in Carpentier and Valko (2014), we consider the quantity $B = (2C'T)^{1/((1+\beta)\alpha)}$ that serves as a cut-off between tail and bulk behaviors. Observe that

$$\left| \mathbb{E}\left[\max_{1 \leq i \leq T} X_i \right] - (TC)^{1/\alpha} \Gamma(1 - 1/\alpha) \right| \leq$$
$$\left| \int_0^\infty \left\{ 1 - \mathbb{P}\left(\max_{1 \leq i \leq T} X_i \leq x \right) - 1 + e^{-TCx^{-\alpha}} \right\} \mathrm{d}x \right|$$
$$\leq \left| \int_0^B \left\{ \mathbb{P}\left(\max_{1 \leq i \leq T} X_i \leq x \right) - e^{-TCx^{-\alpha}} \right\} \mathrm{d}x \right|$$
$$+ \left| \int_B^\infty \left\{ \mathbb{P}\left(\max_{1 \leq i \leq T} X_i \leq x \right) - e^{-TCx^{-\alpha}} \right\} \mathrm{d}x \right|.$$

For $p \in \{2, \beta + 1\}$, we set $D_p = \Gamma(p - \frac{1}{\alpha})/\alpha$. Equipped with this notation, we may write

$$\left| \int_B^\infty \left\{ \mathbb{P}\left(\max_{1 \leq i \leq T} X_i \leq x \right) - e^{-TCx^{-\alpha}} \right\} \mathrm{d}x \right| \leq \frac{4D_2 C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C' D_{\beta+1}}{C^{\beta+1-1/\alpha} T^{\beta-1/\alpha}}.$$

Instead of loosely bounding the bulk term by B, we write

$$\left| \int_0^B \left\{ \mathbb{P}\left(\max_{1 \leq i \leq T} X_i \leq x \right) - e^{-TCx^{-\alpha}} \right\} \mathrm{d}x \right| \leq B \, \mathbb{P}\left(X_1 \leq B \right)^T + \int_0^B e^{-TCx^{-\alpha}} \mathrm{d}x.$$
$$\tag{4}$$

First, using (2) and the inequality $C'B^{-(1+\beta)\alpha} \le CB^{-\alpha}/2$ (a direct consequence of Eq. (3)), we obtain

$$\mathbb{P}(X_1 \le B)^T \le \left(1 - CB^{-\alpha} + C'B^{-(1+\beta)\alpha}\right)^T$$

$$\le \left(1 - \frac{1}{2}CB^{-\alpha}\right)^T \le e^{-\frac{1}{2}TCB^{-\alpha}} = e^{-HT^{\beta/(\beta+1)}}.$$

Second, the integral in Eq. (4) can be bounded as follows:

$$\int_0^B e^{-TCx^{-\alpha}}\mathrm{d}x \le Be^{-TCB^{-\alpha}} = (2C'T)^{1/((1+\beta)\alpha)}e^{-2HT^{\beta/(\beta+1)}}.$$

This concludes the proof.

3 The ExtremeHunter and ExtremeETC algorithms

In this section, the tighter control provided by Theorem 1 is used in order to refine the analysis of the EXTREMEHUNTER algorithm (Algorithm 1) carried out in Carpentier and Valko (2014). This theoretical analysis is also shown to be valid for EXTEMEETC, a novel algorithm we next propose, that greatly improves upon EXTREMEHUNTER, regarding computational efficiency.

3.1 Further Notations and Preliminaries

Throughout the paper, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{1}\{\mathcal{E}\}$ and $\bar{\mathcal{E}}$ means the complementary event of $\mathcal{E}$. We assume that the reward related to each arm $k \in \{1, \ldots, K\}$ is drawn from a $(\alpha_k, \beta_k, C_k, C')$-second order Pareto distribution. Sorting the tail indices by increasing order of magnitude, we use the classical notation for order statistics: $\alpha_{(1)} \le \cdots \le \alpha_{(K)}$. We assume that $\alpha_{(1)} > 1$, so that the random rewards have finite expectations, and suppose that the strict inequality $\alpha_{(1)} < \alpha_{(2)}$ holds true. We also denote by $T_{k,t}$ the number of times the arm k is pulled up to time t. For $1 \le k \le K$ and $i \ge 1$, the r.v. $\widetilde{X}_{k,i}$ is the reward obtained at the i-th draw of arm k if $i \le T_{k,n}$ or a new r.v. drawn from ν_k independent from the other r.v.'s otherwise.

We start with a preliminary lemma supporting the intuition that the tail index α fully governs the extreme bandit problem. It will allow to show next that the algorithm picks the right arm after the exploration phase, see Lemma 2.

Lemma 1. (OPTIMAL ARM) *For n larger than some constant Q_4 depending only on $(\alpha_k, \beta_k, C_k)_{1 \le k \le K}$ and C', the optimal arm for the extreme bandit problem is given by:*

$$k^* = \underset{1 \le k \le K}{\arg\min}\, \alpha_k = \underset{1 \le k \le K}{\arg\max}\, V_k, \tag{5}$$

where $V_k = (nC_k)^{1/\alpha_k}\Gamma(1 - 1/\alpha_k)$.

Proof. We first prove the first equality. It follows from Theorem 1 that there exists a constant Q_2, depending only on $\{(\alpha_k, \beta_k, C_k)\}_{1 \le k \le K}$ and C', such that for any arm $k \in \{1, \ldots, K\}$, $|\mathbb{E}[G_n^{(k)}] - V_k| \le V_k/2$. Then for $k \ne k^*$ we have, for all $n > Q_2$, $V_k/2 \le \mathbb{E}[G_n^{(k)}] \le \mathbb{E}[G_n^{(k^*)}] \le 3V_{k^*}/2$. Recalling that V_k is proportional to n^{1/α_k}, it follows that $\alpha_{k^*} = \min_{1 \le k \le K} \alpha_k$. Now consider the following quantity:

$$Q_3 = \max_{k \ne k^*} \left[\frac{2C_k^{1/\alpha_k} \Gamma(1 - 1/\alpha_k)}{C_{k^*}^{1/\alpha_{k^*}} \Gamma(1 - 1/\alpha_{k^*})} \right]^{1/(1/\alpha_{k^*} - 1/\alpha_k)} . \qquad (6)$$

For $n > Q_4 = \max(Q_2, Q_3)$, we have $V_{k^*} > 2V_k$ for any suboptimal arm $k \ne k^*$, which proves the second equality.

From now on, we assume that n is large enough for Lemma 1 to apply.

3.2 The ExtremeHunter algorithm (Carpentier and Valko, 2014)

Before developing a novel analysis of the extreme bandit problem in Section 3.2 (see Theorem 2), we recall the main features of EXTREMEHUNTER, and in particular the estimators and confidence intervals involved in the indices of this optimistic policy.

Algorithm 1 EXTREMEHUNTER (Carpentier and Valko, 2014)

1: **Input:** K: number of arms, n: time horizon, $b > 0$ such that $b \le \min_{1 \le k \le K} \beta_k$, N: minimum number of pulls of each arm (Eq. (9)).
2: **Initialize:** Pull each arm N times.
3: **for** $k = 1, \ldots, K$ **do**
4:     Compute estimators $\widehat{h}_{k,KN} = \widetilde{h}_k(N)$ (Eq. (8)) and $\widehat{C}_{k,KN} = \widetilde{C}_k(N)$ (Eq. (7))
5:     Compute index $B_{k,KN}$ (Eq. (12))
6: **end for**
7: Pull arm $I_{KN+1} = \arg\max_{1 \le k \le K} B_{k,KN}$
8: **for** $t = KN + 2, \ldots, n$ **do**
9:     Update estimators $\widehat{h}_{I_{t-1}, t-1}$ and $\widehat{C}_{I_{t-1}, t-1}$
10:     Update index $B_{I_{t-1}, t-1}$
11:     Pull arm $I_t = \arg\max_{1 \le k \le K} B_{k, t-1}$
12: **end for**

Theorem 1 states that for any arm $k \in \{1, \ldots, K\}$, $\mathbb{E}[G_n^{(k)}] \approx (C_k n)^{1/\alpha_k} \Gamma(1 - 1/\alpha_k)$. Consequently, the optimal strategy in hindsight always pulls the arm $k^* = \arg\max_{1 \le k \le K} \{(nC_k)^{1/\alpha_k} \Gamma(1 - 1/\alpha_k)\}$. At each round and for each arm $k \in \{1, \ldots, K\}$, EXTREMEHUNTER algorithm (Carpentier and Valko, 2014) estimates the coefficients α_k and C_k (but not β_k, see Remark 2 in Carpentier and Valko (2014)). The corresponding confidence intervals are detailed below. Then, following the *optimism-in-the-face-of-uncertainty* principle (see (Auer

et al., 2002) and references therein), the strategy plays the arm maximizing an optimistic plug-in estimate of $(C_k n)^{1/\alpha_k} \Gamma(1 - 1/\alpha_k)$. To that purpose, Theorem 3.8 in Carpentier and Kim (2014) and Theorem 2 in Carpentier et al. (2014) provide estimators $\widetilde{\alpha}_k(T)$ and $\widetilde{C}_k(T)$ for α_k and C_k respectively, after T draws of arm k. Precisely, the estimate $\widetilde{\alpha}_k(T)$ is given by

$$\widetilde{\alpha}_k(T) = \log\left(\frac{\sum_{t=1}^{T} \mathbb{1}\{X_t > e^r\}}{\sum_{t=1}^{T} \mathbb{1}\{X_t > e^{r+1}\}}\right),$$

where r is chosen in an adaptive fashion based on Lepski's method, see (Lepskiĭ, 1990), while the estimator of C_k considered is

$$\widetilde{C}_k(T) = T^{-2b/(2b+1)} \sum_{i=1}^{T} \mathbb{1}\{\widetilde{X}_{k,i} \geq T^{\widetilde{h}_k(T)/(2b+1)}\}, \tag{7}$$

where

$$\widetilde{h}_k(T) = \min(1/\widetilde{\alpha}_k(T), 1) . \tag{8}$$

The authors also provide finite sample error bounds for $T \geq N$, where

$$N = A_0 (\log n)^{2(2b+1)/b}, \tag{9}$$

with b a known lower bound on the β_k's ($b \leq \min_{1 \leq k \leq K} \beta_k$), and A_0 a constant depending only on $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ and C'. These error bounds naturally define confidence intervals of respective widths Λ_1 and Λ_2 at level δ_0 defined by

$$\delta_0 = n^{-\rho}, \quad \text{where} \quad \rho = \frac{2\alpha_{k^*}}{\alpha_{k^*} - 1}. \tag{10}$$

More precisely, we have

$$\mathbb{P}\left(\left|\frac{1}{\alpha_k} - \widetilde{h}_k(T)\right| \leq \Lambda_1(T), \ \left|C_k - \widetilde{C}_k(T)\right| \leq \Lambda_2(T)\right) \geq 1 - 2\delta_0, \tag{11}$$

where

$$\Lambda_1(T) = D\sqrt{\log(1/\delta_0)} T^{-b/(2b+1)} \text{ and } \Lambda_2(T) = E\sqrt{\log(T/\delta_0)} \log(T) T^{-b/(2b+1)},$$

denoting by D and E some constants depending only on $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ and C'. When $T_{k,t} \geq N$, denote by $\widehat{h}_{k,t} = \widetilde{h}_k(T_{k,t})$ and $\widehat{C}_{k,t} = \widetilde{C}_k(T_{k,t})$ the estimators based on the $T_{k,t}$ observations for simplicity. EXTREMEHUNTER's index $B_{k,t}$ for arm k at time t, the optimistic proxy for $\mathbb{E}[G_n^{(k)}]$, can be then written as

$$B_{k,t} = \widetilde{\Gamma}\left(1 - \widehat{h}_{k,t} - \Lambda_1(T_{k,t})\right)\left(\left(\widehat{C}_{k,t} + \Lambda_2(T_{k,t})\right) n\right)^{\widehat{h}_{k,t} + \Lambda_1(T_{k,t})}, \tag{12}$$

where $\widetilde{\Gamma}(x) = \Gamma(x)$ if $x > 0$ and $+\infty$ otherwise.

On computational complexity. Notice that after the initialization phase, at each time $t > KN$, EXTREMEHUNTER computes estimators $\widehat{h}_{I_t,t}$ and $\widehat{C}_{I_t,t}$, each having a time complexity linear with the number of samples $T_{I_t,t}$ pulled from arm I_t up to time t. Summing on the rounds reveals that EXTREMEHUNTER's time complexity is quadratic with the time horizon n.

3.3 ExtremeETC: a computationally appealing alternative

In order to reduce the restrictive time complexity discussed previously, we now propose the ExtremeETC algorithm, an *Explore-Then-Commit* version of Extreme Hunter, which offers similar theoretical guarantees.

Algorithm 2 ExtremeETC

1: **Input:** K: number of arms, n: time horizon, $b > 0$ such that $b \leq \min_{1 \leq k \leq K} \beta_k$, N: minimum number of pulls of each arm (Eq. (9)).
2: **Initialize:** Pull each arm N times.
3: **for** $k = 1, \ldots, K$ **do**
4:    Compute estimators $\widehat{h}_{k,KN} = \widetilde{h}_k(N)$ (Eq. (8)) and $\widehat{C}_{k,KN} = \widetilde{C}_k(N)$ (Eq. (7))
5:    Compute index $B_{k,KN}$ (Eq. (12))
6: **end for**
7: Set $I_{\text{winner}} = \arg\max_{1 \leq k \leq K} B_{k,KN}$
8: **for** $t = KN + 1, \ldots, n$ **do**
9:    Pull arm I_{winner}
10: **end for**

After the initialization phase, the *winner arm*, which has maximal index $B_{k,KN}$, is fixed and is pulled in all remaining rounds. Then ExtremeETC's time complexity, due to the computation of $\widehat{h}_{k,KN}$ and $\widehat{C}_{k,KN}$ only, is $\mathcal{O}(KN) = \mathcal{O}\left((\log n)^{2(2b+1)/b}\right)$, which is considerably faster than quadratic time achieved by ExtremeHunter. For clarity, Table 1 summarizes time and memory complexities of both algorithms.

Complexity	ExtremeETC	ExtremeHunter
Time	$\mathcal{O}\left((\log n)^{\frac{2(2b+1)}{b}}\right)$	$\mathcal{O}(n^2)$
Memory	$\mathcal{O}\left((\log n)^{\frac{2(2b+1)}{b}}\right)$	$\mathcal{O}(n)$

Table 1: Time and memory complexities required for estimating $(\alpha_k, C_k)_{1 \leq k \leq K}$ in ExtremeETC and ExtremeHunter.

Due to the significant gain of computational time, we used the ExtremeETC algorithm in our simulation study (Section 6) rather than ExtremeHunter.

Controlling the number of suboptimal rounds. We introduce a high probability event that corresponds to the favorable situation where, at each round, all coefficients $(1/\alpha_k, C_k)_{1 \leq k \leq K}$ simultaneously belong to the confidence intervals recalled in the previous subsection.

Definition 1. *The event ξ_1 is the event on which the bounds*

$$\left| \frac{1}{\alpha_k} - \widetilde{h}_k(T) \right| \leq \Lambda_1(T) \quad and \quad \left| C_k - \widetilde{C}_k(T) \right| \leq \Lambda_2(T)$$

hold true for any $1 \leq k \leq K$ and $N \leq T \leq n$.

The union bound combined with (11) yields

$$\mathbb{P}(\xi_1) \geq 1 - 2Kn\delta_0. \tag{13}$$

Lemma 2. *For $n > Q_5$, where Q_5 is the constant defined in (15), EXTREMEETC and EXTREMEHUNTER always pull the optimal arm after the initialization phase on the event ξ_1. Hence, for any suboptimal arm $k \neq k^*$, we have on ξ_1:*

$$T_{k,n} = N \quad and \ thus \quad T_{k^*,n} = n - (K-1)N.$$

Proof. Here we place ourselves on the event ξ_1. For any arm $1 \leq k \leq K$, Lemma 1 in Carpentier and Valko (2014) provides lower and upper bounds for $B_{k,t}$ when $T_{k,t} \geq N$

$$V_k \leq B_{k,t} \leq V_k \left(1 + F \log n \sqrt{\log(n/\delta_0)} T_{k,t}^{-b/(2b+1)} \right), \tag{14}$$

where F is a constant which depends only on $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ and C'. Introduce the horizon Q_5, which depends on $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ and C'

$$Q_5 = \max \left(e^{\left(F\sqrt{1+\rho} A_0^{-b/(2b+1)} \right)^2}, Q_4 \right). \tag{15}$$

Then the following Lemma 3, proved in Appendix A, tells us that for n large enough, the exploration made during the initialization phase is enough to find the optimal arm, with high probability.

Lemma 3. *If $n > Q_5$, we have under the event ξ_1 that for any suboptimal arm $k \neq k^*$ and any time $t > KN$ that $B_{k,t} < B_{k^*,t}$.*

Hence the optimal arm is pulled at any time $t > KN$.

The following result immediately follows from Lemma 2.

Corollary 1. *For n larger than some constant depending only on $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ and C' we have under ξ_1*

$$T_{k^*,n} \geq n/2.$$

Upper bounding the expected extreme regret. The upper bound on the expected extreme regret stated in the theorem below improves upon that given in Carpentier and Valko (2014) for EXTREMEHUNTER. It is also valid for EX-TREMEETC.

Theorem 2. *For* EXTREMEETC *and* EXTREMEHUNTER*, the expected extreme regret is upper bounded as follows*

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

as $n \to +\infty$*. If* $b \geq 1$*, we have in particular* $\mathbb{E}[R_n] = o(1)$ *as* $n \to +\infty$*.*

The proof of Theorem 2 is deferred to Appendix A. It closely follows that of Theorem 2 in Carpentier and Valko (2014), the main difference being that their concentration bound (Theorem 1 therein) can be replaced by our tighter bound (see Theorem 1 in the present paper). Recall that in Theorem 2 in Carpentier and Valko (2014), the upper bound on the expected extreme regret for EXTREMEHUNTER goes to infinity when $n \to +\infty$:

$$\mathbb{E}[R_n] = \mathcal{O}\left(n^{\frac{1}{(1+b)\alpha_{k^*}}}\right). \tag{16}$$

In contrast, in Theorem 2 when $b \geq 1$, the upper bound obtained vanishes when $n \to +\infty$. In the case $b < 1$, the upper bound still improves upon Eq. (16) by a factor $n^{(\alpha_{k^*}b(b+1)-b)/((b+1)\alpha_{k^*})} > n^{b^2/(2\alpha_{k^*})}$.

4 Lower bound on the expected extreme regret

In this section we prove a lower bound on the expected extreme regret for EXTREMEETC and EXTREMEHUNTER in specific cases. We assume now that $\alpha_{(2)} > 2\alpha_{k^*}^2/(\alpha_{k^*} - 1)$ and we start with a preliminary result on second order Pareto distributions, proved in Appendix A.

Lemma 4. *If* X *is a r.v. drawn from a* (α, β, C, C')*-second order Pareto distribution and* r *is a strictly positive constant, the distribution of the r.v.* X^r *is a* $(\alpha/r, \beta, C, C')$*-second order Pareto.*

In order to prove the lower bound on the expected extreme regret, we first establish that the event corresponding to the situation where the highest reward obtained by EXTREMEETC and EXTREMEHUNTER comes from the optimal arm k^* occurs with overwhelming probability. Precisely, we denote by ξ_2 the event such that the bound

$$\max_{k \neq k^*} \max_{1 \leq i \leq N} \widetilde{X}_{k,i} \leq \max_{1 \leq i \leq n-(K-1)N} \widetilde{X}_{k^*,i}.$$

holds true. The following lemma, proved in Appendix A, provides a control of its probability of occurence.

Lemma 5. *For* n *larger than some constant depending only on* $(\alpha_k, \beta_k, C_k)_{1 \leq k \leq K}$ *and* C'*, the following assertions hold true.*

(i) *We have:*
$$\mathbb{P}(\xi_2) \geq 1 - K\delta_0,$$
where δ_0 *is given in Eq. (10).*

(ii) Under the event $\xi_0 = \xi_1 \cap \xi_2$, *the maximum reward obtained by* EX-TREMEETC *and* EXTREMEHUNTER *comes from the optimal arm:*

$$\max_{1 \leq t \leq n} X_{I_t,t} = \max_{1 \leq i \leq n-(K-1)N} \widetilde{X}_{k^*,i}.$$

The following lower bound shows that the upper bound (Theorem 2) is actually tight in the case $b \geq 1$.

Theorem 3. *If* $b \geq 1$ *and* $\alpha_{(2)} > 2\alpha_{k^*}^2/(\alpha_{k^*} - 1)$, *the expected extreme regret of* EXTREMEETC *and* EXTREMEHUNTER *are lower bounded as follows*

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

Proof. Here, π refers to either EXTREMEETC or else EXTREMEHUNTER. In order to bound from below $\mathbb{E}[R_n] = \mathbb{E}[G_n^{(k^*)}] - \mathbb{E}[G_n^{(\pi)}]$, we start with bounding $\mathbb{E}[G_n^{(\pi)}]$ as follows

$$\mathbb{E}\left[G_n^{(\pi)}\right] = \mathbb{E}\left[\max_{1 \leq t \leq n} X_{I_t,t}\right] = \mathbb{E}\left[\max_{1 \leq t \leq n} X_{I_t,t} \mathbb{1}\{\xi_0\}\right] + \mathbb{E}\left[\max_{1 \leq t \leq n} X_{I_t,t} \mathbb{1}\{\bar{\xi}_0\}\right]$$

$$\leq \mathbb{P}(\xi_0)\mathbb{E}\left[\max_{1 \leq t \leq n} X_{I_t,t} \mid \xi_0\right] + \sum_{k=1}^{K} \mathbb{E}\left[\max_{1 \leq i \leq T_{k,n}} \widetilde{X}_{k,i} \mathbb{1}\{\bar{\xi}_0\}\right], \qquad (17)$$

where $\widetilde{X}_{k,i}$ has been defined in Section 3.1. From (ii) in Lemma 5, we have

$$\mathbb{E}\left[\max_{1 \leq t \leq n} X_{I_t,t} \mid \xi_0\right] = \mathbb{E}\left[\max_{1 \leq i \leq n-(K-1)N} \widetilde{X}_{k^*,i} \mid \xi_0\right] . \qquad (18)$$

In addition, in the sum of expectations on the right-hand-side of Eq. (17), $T_{k,n}$ may be roughly bounded from above by n. A straightforward application of Hölder inequality yields

$$\sum_{k=1}^{K} \mathbb{E}\left[\max_{1 \leq i \leq T_{k,n}} \widetilde{X}_{k,i} \mathbb{1}\{\bar{\xi}_0\}\right] \leq \sum_{k=1}^{K} \left(\mathbb{E}\left[\max_{1 \leq i \leq n} \widetilde{X}_{k,i}^{\frac{\alpha_{k^*}+1}{2}}\right]\right)^{\frac{2}{\alpha_{k^*}+1}} \mathbb{P}\left(\bar{\xi}_0\right)^{\frac{\alpha_{k^*}-1}{\alpha_{k^*}+1}} . \qquad (19)$$

From (i) in Lemma 5 and Eq. (13), we have $\mathbb{P}(\bar{\xi}_0) \leq K(2n+1)\delta_0$. By virtue of Lemma 4, the r.v. $\widetilde{X}_{k,i}^{(\alpha_{k^*}+1)/2}$ follows a $(2\alpha_k/(\alpha_{k^*}+1), \beta_k, C_k, C')$-second order Pareto distribution. Then, applying Theorem 1 to the right-hand side of (19) and using the identity (18), the upper bound (17) becomes

$$\mathbb{E}\left[G_n^{(\pi)}\right] \leq \mathbb{E}\left[\max_{1 \leq i \leq n-(K-1)N} \widetilde{X}_{k^*,i} \mathbb{1}\{\xi_0\}\right]$$

$$+ \sum_{k=1}^{K} \left((nC_k)^{\frac{\alpha_{k^*}+1}{2\alpha_k}} \Gamma\left(1 - \frac{\alpha_{k^*}+1}{2\alpha_k}\right) + o\left(n^{\frac{\alpha_{k^*}+1}{2\alpha_k}}\right)\right)^{\frac{2}{\alpha_{k^*}+1}} (K(2n+1)\delta_0)^{\frac{\alpha_{k^*}-1}{\alpha_{k^*}+1}}$$

$$\leq \mathbb{E}\left[\max_{1 \leq i \leq n-(K-1)N} \widetilde{X}_{k^*,i}\right] + \mathcal{O}\left(n^{-(1-1/\alpha_{k^*})}\right), \qquad (20)$$

where the last inequality comes from the definition of δ_0. Combining Theorem 1 and (20) we finally obtain the desired lower bound

$$
\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}\left[G_n^{(k^*)}\right] - \mathbb{E}\left[G_n^{(\pi)}\right] \\
&\geq \Gamma(1 - 1/\alpha_{k^*})C_{k^*}^{1/\alpha_{k^*}}\left(n^{1/\alpha_{k^*}} - (n - (K-1)N)^{1/\alpha_{k^*}}\right) + \mathcal{O}\left(n^{-(1-1/\alpha_{k^*})}\right) \\
&= \frac{\Gamma(1 - 1/\alpha_{k^*})C_{k^*}^{1/\alpha_{k^*}}}{\alpha_{k^*}}(K-1)Nn^{-(1-1/\alpha_{k^*})} + \mathcal{O}\left(n^{-(1-1/\alpha_{k^*})}\right),
\end{aligned}
$$

where we used a Taylor expansion of $x \mapsto (1+x)^{1/\alpha_{k^*}}$ at zero for the last equality.

5 A reduction to classical bandits

The goal of this section is to render explicit the connections between the max K-armed bandit considered in the present paper and a particular instance of the classical Multi-Armed Bandit (MAB) problem.

5.1 MAB setting for extreme rewards

In a situation where only the large rewards matter, an alternative to the max k-armed problem would be to consider the expected cumulative sum of the most 'extreme' rewards, that is, those which exceeds a given high threshold u. For $k \in \{1, \ldots, K\}$ and $t \in \{1, \ldots, n\}$, we denote by $Y_{k,t}$ these new rewards

$$
Y_{k,t} = X_{k,t}\mathbb{1}\{X_{k,t} > u\} .
$$

In this context, the classical MAB problem consists in maximizing the expected cumulative gain

$$
\mathbb{E}\left[G^{\mathrm{MAB}}\right] = \mathbb{E}\left[\sum_{t=1}^{n} Y_{I_t,t}\right].
$$

It turns out that for a high enough threshold u, the unique optimal arm for this MAB problem, $\arg\max_{1 \leq k \leq K} \mathbb{E}[Y_{k,1}]$, is also the optimal arm k^* for the max k-armed problem. We still assume second order Pareto distributions for the random variables $X_{k,t}$ and that all the hypothesis listed in Section 3.1 hold true. The rewards $\{Y_{k,t}\}_{1 \leq k \leq K, 1 \leq t \leq T}$ are also heavy-tailed so that it is legitimate to attack this MAB problem with the ROBUST UCB algorithm (Bubeck et al., 2013), which assumes that the rewards have finite moments of order $1 + \epsilon$

$$
\max_{1 \leq k \leq K} \mathbb{E}\left[|Y_{k,1}|^{1+\epsilon}\right] \leq v, \tag{21}
$$

where $\epsilon \in (0, 1]$ and $v > 0$ are known constants. Given our second order Pareto assumptions, it follows that Eq. (21) holds with $1 + \epsilon < \alpha_{(1)}$. Even if the knowledge of such constants ϵ and v is a strong assumption, it is still fair to compare ROBUST UCB to EXTREMEETC/HUNTER, which also has strong requirements. Indeed, EXTREMEETC/HUNTER assumes that b and n are known and

verify conditions depending on unknown problem parameters (e.g. $n \geq Q_1$, see Eq. (3)).

The following Lemma, whose the proof is postponed to Appendix A, ensures that the two bandit problems are equivalent for high thresholds.

Lemma 6.

$$\text{If} \quad u > \max\left(1, \left(\frac{2C'}{\min_{1 \leq k \leq K} C_k}\right)^{\frac{1}{\min_{1 \leq k \leq K} \beta_k}}, \left(\frac{3 \max_{1 \leq k \leq K} C_k}{\min_{1 \leq k \leq K} C_k}\right)^{\frac{1}{\alpha_{(2)} - \alpha_{(1)}}}\right),$$
(22)

then the unique best arm for the MAB problem is $\arg\min_{1 \leq k \leq K} \alpha_k = k^*$.

Remark 1. Tuning the threshold u based on the data is a difficult question, outside our scope. A standard practice is to monitor a relevant output (e.g. estimate of α) as a function of the threshold u and to pick the latter as low as possible in the stability region of the output. This is related to the Lepski's method, see e.g. Boucheron and Thomas (2015), Carpentier and Kim (2014), Hall and Welsh (1985).

5.2 Robust UCB algorithm (Bubeck et al., 2013)

For the sake of completeness, we recall below the main feature of ROBUST UCB and make explicit its theoretical guarantees in our setting. The bound stated in the following proposition is a direct consequence of the regret analysis conducted by Bubeck et al. (2013).

Proposition 1. *Applying the* ROBUST UCB *algorithm of (Bubeck et al., 2013) to our MAB problem, the expected number of times we pull any suboptimal arm* $k \neq k^*$ *is upper bounded as follows*

$$\mathbb{E}[T_{k,n}] = \mathcal{O}\left(\log n\right) .$$

Proof. See proof of Proposition 1 in Bubeck et al. (2013).

Hence, in expectation, ROBUST UCB pulls fewer times suboptimal arms than EXTREMEETC/HUNTER. Indeed with EXTREMEETC/HUNTER, $T_{k,n} \geq N = \Theta((\log n)^{2(2b+1)/b})$.

Remark 2. Proposition 1 may be an indication that the Robust UCB approach performs better than EXTREMEETC/HUNTER. Nevertheless, guarantees on its expected extreme regret require sharp concentration bounds on $T_{k,n}$ $(k \neq k^*)$, which is out of the scope of this paper and left for future work.

Algorithm 3 ROBUST UCB with truncated mean estimator (Bubeck et al., 2013)

1: **Input:** $u > 0$ s.t. Eq. (22), $\epsilon \in (0, 1]$ and $v > 0$ s.t. Eq. (21).
2: **Initialize:** Pull each arm once.
3: **for** $t \geq K + 1$ **do**
4:    **for** $k = 1, \ldots, K$ **do**
5:       Update truncated mean estimator
     $\widehat{\mu_k} \leftarrow \frac{1}{T_{k,t-1}} \sum_{s=1}^{t-1} Y_{k,s} \mathbb{1}\left\{ I_s = k, Y_{k,s} \leq \left(\frac{v T_{k,s}}{\log(t^2)} \right)^{\frac{1}{1+\epsilon}} \right\}$
6:       Update index
     $B_k \leftarrow \widehat{\mu_k} + 4 v^{1/(1+\epsilon)} \left(\frac{\log t^2}{T_{k,t-1}} \right)^{\epsilon/(1+\epsilon)}$
7:    **end for**
8:    Play arm $I_t = \arg\max_{1 \leq k \leq K} B_k$
9: **end for**

---

6   Numerical experiments

In order to illustrate some aspects of the theoretical results presented previously, we consider a time horizon $n = 10^5$ with $K = 3$ arms and exact Pareto distributions with parameters given in Table 2. Here, the optimal arm is the second one (incidentally, the distribution with highest mean is the first one).

| | Arms | | |
	$k = 1$	$k^* = 2$	$k = 3$
α_k	15	1.5	10
C_k	10^8	1	10^5
$\mathbb{E}[X_{k,1}]$	3.7	3	3.5
$\mathbb{E}[\max_{1 \leq t \leq n} X_{k,t}]$	7.7	$5.8 \cdot 10^3$	11

Table 2: Pareto distributions used in the experiments.

We have implemented ROBUST UCB with parameters $\epsilon = 0.4$, which satisfies $1 + \epsilon < \alpha_2 = 1.5$, v achieving the equality in Eq. (21) (ideal case) and a threshold u equal to the lower bound in Eq. (22) plus 1 to respect the strict inequality. EXTREMEETC is runned with $b = 1 < +\infty = \min_{1 \leq k \leq K} \beta_k$. In this setting, the most restrictive condition on the time horizon, $n > KN \approx 7000$ (given by Eq. (9)), is checked, which places us in the validity framework of EXTREMEETC. The resulting strategies are compared to each other and to the random strategy pulling each arm uniformly at random, but not to THRESHOLD ASCENT algorithm (Streeter and Smith, 2006) which is designed only for bounded rewards. Precisely, 1000 simulations have been run and Figure 1 depicts the extreme regret (1) in each setting averaged over these 1000 trajectories. These experiments
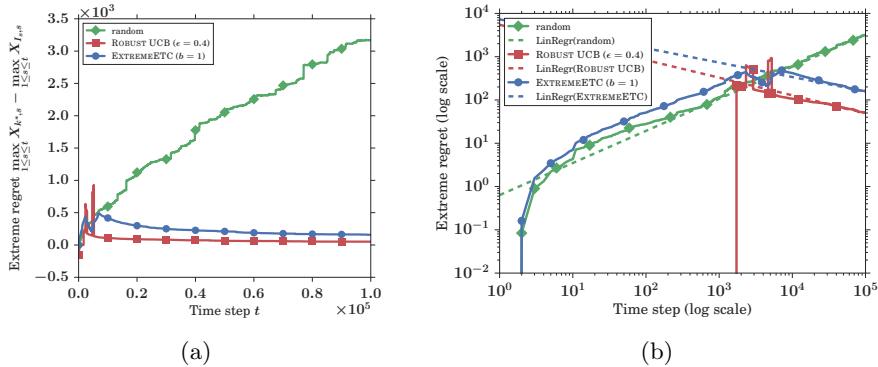
Fig. 1: Averaged extreme regret (over 1000 independent simulations) for Ex-tremeETC, Robust UCB and a uniformly random strategy. Fig. 1b is the log-log scaled counterpart of Fig. 1a with linear regressions computed over $t = 5 \cdot 10^4, \ldots, 10^5$.

empirically support the theoretical bounds in Theorem 2: the expected extreme regret of ExtremeETC converges to zero for large horizons. On the log-log scale (Fig. 1b), ExtremeETC's extreme regret starts linearly decreasing after the initialization phase, at $n > KN \approx 7000$, which is consistent with Lemma 2. The corresponding linear regression reveals a slope ≈ -0.333 (with a coefficient of determination $R^2 \approx 0.97$), which confirms Theorem 2 and Theorem 3 yielding the theoretical slope $-(1 - 1/\alpha_{k^*}) = -1/3$.

7 Conclusion

This paper brings two main contributions. It first provides a refined regret bound analysis of the performance of the ExtremeHunter algorithm in the context of the max K-armed bandit problem that significantly improves upon the results obtained in the seminal contribution Carpentier and Valko (2014), also proved to be valid for ExtremeETC, a computationally appealing alternative we intro-duce. In particular, the obtained upper bound on the regret converges to zero for large horizons and is shown to be tight when the tail of the rewards is sufficiently close to a Pareto tail (second order parameter $b \geq 1$). On the other hand, this paper offers a novel view of this approach, interpreted here as a specific version of a classical solution (*Robust UCB*) of the MAB problem, in the situation when only very large rewards matter.

Based on these encouraging results, several lines of further research can be sketched. In particular, future work will investigate to which extent the lower bound established for ExtremeETC/Hunter holds true for any strategy with exploration stage of the same duration, and whether improved performance is achievable with alternative stopping criteria for the exploration stage.

Acknowledgments

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Boucheron, S. and Thomas, M. (2015). Tail index estimation, concentration and adaptivity. *Electron. J. Statist.*, 9(2):2751–2792.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Carpentier, A. and Kim, A. K. (2014). Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, 25:1133–1144.

Carpentier, A., Kim, A. K., et al. (2014). Honest and adaptive confidence interval for the tail coefficient in the pareto model. *Electronic Journal of Statistics*, 8(2):2066–2110.

Carpentier, A. and Valko, M. (2014). Extreme bandits. In *Advances in Neural Information Processing Systems 27*, pages 1089–1097. Curran Associates, Inc.

Carpentier, A. and Valko, M. (2015). Simple regret for infinitely many armed bandits. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1133–1141.

Cicirello, V. A. and Smith, S. F. (2005). The max k-armed bandit: A new model of exploration applied to search heuristic selection. In *The Proceedings of the Twentieth National Conference on Artificial Intelligence*, volume 3, pages 1355–1361. AAAI Press.

David, Y. and Shimkin, N. (2016). Pac lower bounds and efficient algorithms for the max k-armed bandit problem. In *Proceedings of The 33nd International Conference on Machine Learning*.

Hall, P. and Welsh, A. H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13(1):331–341.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

Nishihara, R., Lopez-Paz, D., and Bottou, L. (2016). No regret bound for extreme bandits. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Number vol. 10 in Heavy-tail Phenomena: Probabilistic and Statistical Modeling. Springer.

Streeter, M. J. and Smith, S. F. (2006). A simple distribution-free approach to the max k-armed bandit problem. In *International Conference on Principles and Practice of Constraint Programming*, pages 560–574. Springer.