Improving the Pareto UCB1 Algorithm on the Multi-Objective Multi-Armed Bandit

Audrey Durand[*] **Charles Bordet**[*] **Christian Gagné**[†]
Université Laval

Abstract

In this work, we introduce a straightforward approach for bounding the regret of Multi-Objective Multi-Armed Bandit (MO-MAB) heuristics extended from standard bandit algorithms. The proposed methodology allows us to easily build upon the regret analysis of the heuristics in the standard bandit setting. Using our approach, we improve the Pareto UCB1 algorithm, that is the multi-objective extension of the seminal UCB1, by performing a tighter regret analysis. The resulting Pareto UCB1* also has the advantage of being empirically usable without any approximation.

1 Multi-Objective Multi-Armed Bandit

The Multi-Objective Multi-Armed Bandit (MO-MAB) setting [1] is described by a set of arms $\mathcal{K}$ associated with a set of random variables vectors $\{\boldsymbol{x}_{k,t}|t \geq 1\}$ for each k in $\mathcal{K}$. Let N be the number of objectives. Vector $\boldsymbol{x}_{k,t} = [x_{k,t,1}, \ldots, x_{k,t,N}]$ indicates the random outcome of the k-th arm in its t-th trial, where $x_{k,t,i} \in \mathbb{R}$. We consider the stochastic setting where all $\boldsymbol{x}_{k,t}$ associated with k are independent and identically distributed according to some unknown distribution with unknown expectation vector $\boldsymbol{\mu}_k = [\mu_{k,1}, \ldots, \mu_{k,N}]$.

Given two arms a and b, a is said to dominate, or Pareto-dominate, b (denoted $a \succeq b$) if $\mu_{a,i} \geq \mu_{b,i}$ for every objective i. The dominance is strict (denoted $a \succ b$) if $\mu_{a,i} > \mu_{b,i}$ for every objective i. Finally, the two arms are incomparable (denoted $a \parallel b$) if $a \not\succ b$ and $b \not\succ a$. The set of optimal arms contains all the non-dominated arms such that $\mathcal{K}^* = \{k \in \mathcal{K} | \nexists k' \in \mathcal{K}, \boldsymbol{\mu}_{k'} \succ \boldsymbol{\mu}_k\}$. The Pareto front $\mathcal{P}$, also referred to as the Pareto-optimal set, contains the expectations of the optimal arms such that $\mathcal{P} = \{\boldsymbol{\mu}_{k^*} \ \forall k^* \in \mathcal{K}^*\}$. In this work, we consider the setting where all optimal arms are considered equivalent, that is we are not biased in playing any of them more than others (in $\mathcal{K}^*$).

The problem can be formulated as a game where a player sequentially selects arms in $\mathcal{K}$ and observes rewards according to the played arms. Let $k(t)$ denote the arm played at episode t and the reward $\boldsymbol{r}(t) = \boldsymbol{x}_{k(t),t}$. The goal is to simultaneously maximize the reward over time for all objectives. Therefore, we want to play as much as possible any optimal arm in $\mathcal{K}^*$. Let $n_k(t)$ denote the number of times arm k has been played of to time $t-1$. The performance is measured with the expected regret metric denoted as

$$\mathbb{E}[R(T)] = \sum_{k \in \mathcal{K}} \mathbb{E}[n_k(T)]\Delta_k, \tag{1}$$

where T is the number of episodes performed up to now and Δ_k corresponds to the regret of playing arm k instead of an optimal arm (in $\mathcal{K}^*$).

A typical approach for adapting standard bandit heuristics to the MO-MAB setting relies on the concept of Pareto-dominance. Instead of playing the arm that maximizes the expected regret, one

[*]{audrey.durand.2, charles.bordet.1}@ulaval.ca
[†]christian.gagne@gel.ulaval.ca

might randomly play an arm among those for which the expected regret vector is non-dominated, that would be the candidate set $\mathcal{C}(t)$. The usual approach for computing the regret bounds for bandit heuristics consists in bounding $\mathbb{E}[n_k(T)]$. In their analysis of UCB1, UCB2, and ε_n-greedy, Auer et al. [2] bound the number of times that the expected value of each suboptimal arm k could be higher than the expected value of the optimal arm. Working with the Pareto-dominance in the MO-MAB setting, the analysis can be done by bounding the number of times that the expected vector value of each suboptimal arm k happens to be in the candidate set $\mathcal{C}(t)$.

2 Pareto UCB1

Let $\hat{\boldsymbol{\mu}}_k(t)$ denote the mean of the observed rewards $\{\boldsymbol{r}(\tau)|k(\tau)=k, \tau=1,\ldots,t-1\}$. Drugan and Nowe [1] have extended the seminal UCB1 [2] heuristic to the MO-MAB setting. In the resulting Pareto UCB1 algorithm, a multi-objective expected upper bound

$$\boldsymbol{u}_k(t) = \hat{\boldsymbol{\mu}}_k(t) + \left[\sqrt{\frac{2\ln t \sqrt[4]{N|\mathcal{K}^*|}}{n_k(t)}}\right]^N$$

is computed for each arm k, where N is the number of objectives. A candidate set $\mathcal{C}(t)$ is then built, containing all arms for which the upper bound is non-dominated, and $k(t)$ is uniformly picked in $\mathcal{C}(t)$. However, because $\mathcal{K}^*$ is typically unknown, the authors provide the following upper confidence bound for empirical use:

$$\boldsymbol{u}_k(t) = \hat{\boldsymbol{\mu}}_k(t) + \left[\sqrt{\frac{2\ln t \sqrt[4]{N|\mathcal{K}|}}{n_k(t)}}\right]^N .$$

In their regret analysis, Drugan and Nowe [1] bound $\mathbb{E}[n_k(T)]$ for a given suboptimal arm k by the number of times that $\boldsymbol{u}_k(t)$ is non-dominated by $\boldsymbol{u}_{k^*}(t)$, for all episodes t until T, for all k^* in $\mathcal{K}^*$, such that

$$\begin{aligned}
\mathbb{E}[n_k(T)] &= 1 + \sum_{t=K+1}^{T} I(k(t)=k) \\
&\leq l + \sum_{t=K+1}^{T} I(k(t)=k, n_k(t) \geq l) \\
&\leq l + \sum_{t=K+1}^{T} \sum_{k^*\in\mathcal{K}^*} I(\boldsymbol{u}_{k^*}(t) \nsucc \boldsymbol{u}_k(t)),
\end{aligned}$$

where $l > 0$ is an arbitrary number, leading to the regret upper bound

$$\mathbb{E}[R(T)] \leq \sum_{k\notin\mathcal{K}^*} \frac{8\ln\left(T\sqrt[4]{N|\mathcal{K}^*|}\right)}{\Delta_k} + \left(1+\frac{\pi^2}{3}\right)\sum_{k\notin\mathcal{K}^*}\Delta_k.$$

In their analysis, the non-dominance of $\boldsymbol{u}_k(t)$ by $\boldsymbol{u}_{k^*}(t)$ for several k^* in the same episode are counted as multiple plays of arm k, which explains the dependence on the size of the optimal set in the regret upper bound.

3 Pareto UCB1*

In this work, we use the concept of ε-dominance [3] to assign an optimal arm k^* to each suboptimal arm k, such that k^* is the one that dominates k the most. Given two arms a and b, a is said to ε-dominate b (denoted $a \succ_\varepsilon b$) if $\mu_{a,i} + \varepsilon \geq \mu_{b,i}$ for every objective i and the strict inequality is true for at least one objective. It corresponds to the smallest value which must be added to every objective so that the resulting vector is not strictly dominated by any member of the Pareto-optimal set. In other words, it measures how far a suboptimal arm is from belonging to $\mathcal{P}$. Therefore, we

Algorithm 1 Pareto UCB1*

assume N is the number of objectives
maintain vector $\hat{\boldsymbol{\mu}}_k(t)$ as the empirical mean of the observed rewards for each arm k up to time $t-1$.
$t = 0$
loop
  $t = t + 1$
  for all arms k in $\mathcal{K}$ **do**
    $\boldsymbol{u}_k(t) = \hat{\boldsymbol{\mu}}_k(t) + \left[\sqrt{\frac{2 \ln t \sqrt[4]{N}}{n_k(t)}}\right]^N$
  end for
  $\mathcal{C}(t) = \{k | \nexists k' \in \mathcal{K}, \hat{\boldsymbol{\mu}}_{k'} \succ \hat{\boldsymbol{\mu}}_k\}$
  randomly select arm $k(t)$ in $\mathcal{C}(t)$
  play $k(t)$, observe $\boldsymbol{r}(t)$, and update $\hat{\boldsymbol{\mu}}_{k(t)}(t)$
end loop

assign to each suboptimal arm k the optimal arm k^* which maximizes the ε-dominance. Arm k^* is therefore the optimal arm with the most chances of dominating arm k. We then bound the number of times that arm k is played by bounding the number of times that its expected vector value is non-dominated by the expected value of k^*.

Using the proposed approach, we improve Pareto UCB1 to obtain a tighter bound on the regret that does not depend on the size of the optimal set $\mathcal{K}^*$. This is also the case for the upper confidence bound in the resulting Pareto UCB1* given by Algorithm 1, which removes the problem encountered by Drugan and Nowe [1] when computing $\boldsymbol{u}_k(t)$ empirically.

Theorem 1. *For the K-armed stochastic multi-objective bandit problem with rewards in $[0,1]^N$, Pareto UCB1* has expected regret*

$$\mathbb{E}[R(T)] \leq \sum_{k \notin \mathcal{K}^*} \frac{8 \ln (T \sqrt[4]{N})}{\Delta_k} + \left(1 + \frac{\pi^2}{3}\right) \sum_{k \notin \mathcal{K}^*} \Delta_k,$$

in time T.

Fact 1 (N-dimensional Chernoff-Hœffding bound)**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$ be independent N-dimensional random variables sampled with $\mathbb{E}[\boldsymbol{X}_m] = [p_{m,1}, \ldots, p_{m,N}]$ (not necessarily equal), $\bar{\boldsymbol{X}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{X}_m$, and $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \frac{1}{M} \sum_{m=1}^{M} [p_{m,1}, \ldots, p_{m,N}]$.*

We consider the following generalization of the standard Chernoff-Hœffding bound for N-dimensional spaces provided by Drugan et al. [1]

$$\mathbb{P}[\bar{\boldsymbol{X}} \not\prec \boldsymbol{\mu} + [\lambda]^N] \leq N e^{-2M\lambda^2} \quad \textit{and} \quad \mathbb{P}[\bar{\boldsymbol{X}} \not\succ \boldsymbol{\mu} - [\lambda]^N] \leq N e^{-2M\lambda^2},$$

where $\lambda \geq 0$.

Definition 1 ($k(t)$, $\boldsymbol{u}_k(t)$, $\mathcal{C}(t)$)**.** *Let $k(t)$ denote the arm played at time t. On each episode t, an upper confidence bound $\boldsymbol{u}_k(t)$ is computed for each arm k. All arms k in $\mathcal{K}$ for which $\boldsymbol{u}_k(t)$ is non-dominated constitute the candidates set $\mathcal{C}(t)$. The arm $k(t)$ is uniformly selected among members of $\mathcal{C}(t)$.*

Definition 2 (Quantities $n_k(t)$, $\hat{\boldsymbol{\mu}}_k(t)$)**.** *Let $n_k(t)$ denote the number of plays of arm k until time $t-1$. Empirical mean $\hat{\boldsymbol{\mu}}_k(t) = [\hat{\mu}_{k,1}(t), \ldots, \hat{\mu}_{k,N}(t)]$ is defined as $\hat{\mu}_{k,i}(t) = \left(\sum_{\tau=1:k(\tau)=k}^{t-1} r_k(\tau)\right) / \left(n_k(t) + 1\right)$. Note that $\hat{\mu}_{k,i}(t) = 0$ when $n_k(t) = 0$.*

The proof of Theorem 1 follows the proof for Pareto UCB1 [1] and UCB1 [2]. Let $l > 0$ be an arbitrary number, we denote $c_n(t) = \sqrt{\frac{2 \ln (t \sqrt[4]{N})}{n}}$. We can bound the expected number of plays of

a suboptimal arm k as follows:

$$\begin{aligned}
\mathbb{E}[n_k(T)] &= 1 + \sum_{t=K+1}^{T} I(k(t)=k) \\
&\leq l + \sum_{t=K+1}^{T} I(k(t)=k, n_k(t) \geq l) \\
&\leq l + \sum_{t=K+1}^{T} I(\boldsymbol{u}_{k^*}(t) \nsucc \boldsymbol{u}_k(t), n_k(t) \geq l) \\
&\leq l + \sum_{t=K+1}^{T} I(\hat{\boldsymbol{\mu}}_{k^*}(t) + [c_{n^*_k(t)}(t-1)]^N \nsucc \hat{\boldsymbol{\mu}}_k(t) + [c_{n_k(t)}(t-1)]^N, n_k(t) \geq l) \\
&\leq l + \sum_{t=K+1}^{T} I(\min_{0<s<t} \hat{\boldsymbol{\mu}}_{k^*}(s) + [c_s(t-1)]^N \nsucc \max_{l \leq s_k < t} \hat{\boldsymbol{\mu}}_k(s_k) + [c_{s_k}(t-1)]^N) \\
&\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_k=1}^{t-1} I(\hat{\boldsymbol{\mu}}_{k^*}(s) + [c_s(t)]^N \nsucc \hat{\boldsymbol{\mu}}_k(s_k) + [c_{s_k}(t)]^N),
\end{aligned} \tag{2}$$

where k^* is the optimal arm that dominates k the most. The non-domination in the last inequality implies that at least one of the following conditions must hold: $\hat{\boldsymbol{\mu}}_{k^*}(s) \nsucc \boldsymbol{\mu}_{k^*} + [c_s(t)]^N$, $\hat{\boldsymbol{\mu}}_k(s_k) \nprec \boldsymbol{\mu}_k + [c_{s_k}(t)]^N$, and $\boldsymbol{\mu}_{k^*} \nsucc \boldsymbol{\mu}_k + 2 \cdot [c_{s_k}(t)]^N$. We bound the probability of the first two events using Fact 1 such that

$$\mathbb{P}[\hat{\boldsymbol{\mu}}_{k^*}(s) \nsucc \boldsymbol{\mu}_{k^*} + [c_s(t)]^N] \leq t^{-4} \quad \text{and} \tag{3}$$

$$\mathbb{P}[\hat{\boldsymbol{\mu}}_k(s_k) \nprec \boldsymbol{\mu}_k + [c_{s_k}(t)]^N] \leq t^{-4}. \tag{4}$$

For $s_k \geq \frac{8 \ln t \sqrt[4]{N}}{\Delta_k^2}$, we have

$$\mu_{k^*,i} - \mu_{k,i} - 2c_{s_k}(t) \geq \mu_{k^*,i} - \mu_{k,i} - \Delta_k \geq 0$$

for at least one objective i. Therefore, by using $l = \left\lceil \frac{8 \ln t \sqrt[4]{N}}{\Delta_k^2} \right\rceil$ along with Equations 3 and 4 in Equation 2, we obtain

$$\mathbb{E}[n_k(T)] \leq \left\lceil \frac{8 \ln t \sqrt[4]{N}}{\Delta_k^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t} \sum_{s_k=1}^{t} 2t^{-4} \leq \frac{8 \ln t \sqrt[4]{N}}{\Delta_k^2} + 1 + \frac{\pi^2}{3},$$

which leads to Theorem 1 when substituted into Equation 1.

4 Conclusion

In this work, we have introduced a conceptually simple approach for performing the regret analysis of MO-MAB algorithms based on the most dominant optimal arms instead of all the optimal set. Using our methodology, we have proposed an improved version of the Pareto UCB1 heuristic. Pareto UCB1* has tighter regret bounds and it can be used empirically without approximations as we have removed the inconvenient dependence of Pareto UCB1 upon the size of the optimal set.

References

[1] M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2013.

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

[3] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary computation*, 10(3):263–82, 2002.