Pareto Regret Analyses in Multi-objective Multi-armed Bandit

Mengfan Xu [1] **Diego Klabjan** [1]

Abstract

We study Pareto optimality in multi-objective multi-armed bandit by providing a formulation of adversarial multi-objective multi-armed bandit and defining its Pareto regrets that can be applied to both stochastic and adversarial settings. The regrets do not rely on any scalarization functions and reflect Pareto optimality compared to scalarized regrets. We also present new algorithms assuming both with and without prior information of the multi-objective multi-armed bandit setting. The algorithms are shown optimal in adversarial settings and nearly optimal up to a logarithmic factor in stochastic settings simultaneously by our established upper bounds and lower bounds on Pareto regrets. Moreover, the lower bound analyses show that the new regrets are consistent with the existing Pareto regret for stochastic settings and extend an adversarial attack mechanism from bandit to the multi-objective one.

1. Introduction

Multi-armed bandit (MAB) is a sequential paradigm where players choose arms and receive reward values from an environment at each time step. It usually aims at maximizing the cumulative reward of the player, or equivalently minimizing the regret formulated as the difference between the rewards of the best arm and the obtained rewards of the player. Multi-objective Multi-armed bandit (MO-MAB) extends reward values of arms to multi-dimensional reward vectors and thereby changing the nature of the MAB problem significantly as a result of the complicated order relationships of vectors. Rewards can be optimal in one dimension but sub-optimal in other dimensions, leading to multiple optimal arms. Formally, given time horizon T, at time step $t \leq T$ the player chooses one arm a_t among K arms and receives the D-dimensional vector $r^{a_t,t}$ among rewards $(r^{i,t})_{i=1}^K$. The goal is to optimize the total reward vector $\sum_{t=1}^T r^{a_t,t}$ by minimizing some regret metric measuring how far the player is away from optimality.

There are two ways to define optimality: Pareto optimality in the reward vector space and Scalarized optimality by scalarizing reward vectors. Pareto optimality admits a Pareto optimal front defined as the set of rewards of optimal arms determined by the Pareto order relationship. With limited information based on the definition of MO-MAB, it is a great challenge to directly estimate the Pareto optimal front while crucial for an optimal strategy. Scalarization is often used to transform reward vectors to scalars and thereby reducing MO-MAB to MAB depending on scalarization functions, which, however, has the following several limitations. First, how to properly choose such scalarization functions may be troublesome. Linear and Chebyshev options have been discussed in (Drugan and Nowe, 2013), despite of the fact that they may not fully explore non-convex Pareto optimal fronts. Secondly, algorithms can be less robust since their performance highly depends on the specific scalarization functions. In other words, what is optimal for a scalarization function, can result in large scalarized regrets for other functions. Therefore, Pareto optimality in the reward vector space makes more sense and thereby being the main focus herein.

MAB usually falls into the category of either stochastic MAB or adversarial MAB depending on how rewards are generated. Following similar assumptions on the reward process, we consider MO-MAB as stochastic MO-MAB and adversarial MO-MAB, respectively. In stochastic MO-MAB, the rewards of each arm at different time steps are assumed to be i.i.d., following a multi-dimensional distribution of which the mean vector is time-invariant. Henceforth, the Pareto optimal set is the set of arms with mean vectors that are Pareto optimal and the Pareto optimal reward set (front) is the set of corresponding mean vectors, which are constant over time. A regret metric concerning Pareto optimality, namely Pareto regret, is then defined as the cumulative distance between the expected reward received at each time step and the Pareto optimal front (Drugan and Nowe, 2013). However, this regret measure cannot generalize to adversarial MO-MAB by noting that the Pareto optimal front is ill-defined. Precisely, for adversarial MO-MAB, the rewards are arbitrarily specified by adversaries at each time

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, U.S.A.. Correspondence to: Mengfan Xu <MengfanXu2023@u.northwestern.edu>, Diego Klabjan <d-klabjan@northwestern.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

step and thereby varying with time and even depending on past observations. Meanwhile, when the adversary generates rewards based on a distribution depending on the state (context) at each time step, i.e. a context-dependent reward generator, it boils down to contextual MO-MAB. It necessitates a new way to determine Pareto optimality, Pareto optimal front and subsequently a Pareto regret measure in adversarial settings, which has not yet been studied.

Traditionally, algorithms designed for MO-MAB work either by modifying the methods for MAB (MAB-based) or by importing the mechanisms from Multi-objective Optimization (MOO) (MOO-based). More specifically, MAB-based algorithms alternate the estimations of the best arm by approximating a Pareto optimal set, from which an arm is randomly selected. Examples include Pareto UCB (Drugan and Nowe, 2013) as optimistic in face of uncertainty and MOSS++ (Zhu and Nowak, 2020) as an extension of MOSS (Audibert and Bubeck, 2009). Their formulations and analyses are limited to stochastic MO-MAB. While a lot of attention has been given to algorithms for variants of stochastic MO-MAB (see (Hüyük and Tekin, 2021; Van Moffaert et al., 2014)), a counterpart remains unexplored in adversarial settings, including the lack of a valid formulation and analyses which we fully cover herein. Moreover, the aforementioned MAB-based algorithms may result in linear Pareto regret when applied to our defined adversarial settings. Precisely, we provide a regret analysis of Pareto UCB in the adversarial regime where Pareto UCB suggests a linear regret.

Some MOO-based approaches adapt Multi-objective Evolutionary Algorithms (Hong et al., 2021) to bandit problems and (Yahyaa et al., 2014) propose the annealing knowledge gradient descent method, both of which, however, are examined empirically without theoretical guarantees. Analyses for MOO under uncertainty provide possibilities for adversarial MO-MAB given that the objective functions can be black-box. Recently, a minimax regret formulation has been suggested by (Groetzner and Werner, 2022), though Pareto optimality is again neither defined nor potentially guaranteed by proposing algorithms with an optimal minimax regret. An algorithm that guarantees theoretically optimal Pareto regret for adversarial MO-MAB has not yet been proposed, let alone achieving optimality for both adversarial and stochastic MO-MAB.

Herein we propose a formulation of the adversarial MO-MAB problem where rewards are determined arbitrarily under no assumptions on their distributions. Moreover, we formally introduce its Pareto optimal front and Pareto regret and Pareto pseudo regret, which makes it a possibility to consider algorithms from a perspective of Pareto optimality. To our best knowledge, this is the first work formally introducing the general MO-MAB setting and considering the task of Pareto regret optimization in adversarial MO-MAB, in line with both MAB and stochastic MO-MAB. Surprisingly, we can extend these regrets to stochastic settings without any modifications. On the basis of existing literature on stochastic MO-MAB, the achievements of this paper are to build the theoretical connection between general MO-MAB and MAB which allows us to characterize Pareto regret generally, and to develop theoretically effective methods that can achieve Pareto optimality in both stochastic and adversarial MO-MAB.

In this paper, we characterize the Pareto regret explicitly and propose algorithms that achieve regrets of order $\sqrt{T}$ in adversarial MO-MAB and regrets of order $\log T$ or $(\log T)^2$ in stochastic MO-MAB. More specifically, when given whether it is a stochastic schema or adversarial schema before the game starts, our algorithm MO-KS, which switches between the UCB algorithm and the EXP3.P algorithm, obtains Pareto regret of the same orders as the regret in MAB in the sense that the regret is of order $\log T$ and $\sqrt{T}$ in stochastic MO-MAB and adversarial MO-MAB, respectively. Otherwise, when this prior information is not available a different algorithm MO-US is proposed herein that combines the EXP3 and UCB algorithms, as a modification of the existing IEXP3++ algorithm for MAB (Seldin and Lugosi, 2017). Consequently, its Pareto pseudo regret has an order of $(\log T)^2$ in stochastic MO-MAB and $\sqrt{T}$ in adversarial MO-MAB, respectively, without necessarily knowing the settings of MO-MAB.

To claim optimality of our proposed algorithms and to provide the worst-case scenario analyses for the proposed framework, we are the first to provide regret lower bound analyses for both stochastic and adversarial MO-MAB. More precisely, we show that in stochastic MO-MAB, the expected Pareto regret and the Pareto pseudo regret cannot be smaller than order $\log T$ whatever algorithms are deployed. For adversarial MO-MAB, a regret of order greater than or equal to $\sqrt{T}$ results from any randomized algorithm. The lower bounds on Pareto regret match the upper bounds of MO-KS in that they have the same order with respect to T, which implies that it is optimal in both settings. Note that for MO-US, the lower bound coincides with the upper bound in adversarial settings and is almost the same as the upper bound in stochastic ones up to the $\log T$ multiplicative factor, i.e. MO-US is optimal in adversarial MO-MAB and nearly optimal up to a $\log T$ factor in stochastic MO-MAB. Moreover, we derive the lower bound of Pareto UCB, a specific algorithm that is optimal in stochastic MO-MAB, by constructing instances in adversarial MO-MAB. We report that Pareto UCB can result in a linear Pareto regret far away from optimal, which further necessitates the proposed algorithms. As a by-product, the instances form a generalization of an adversarial attack on UCB in MAB (Jun et al., 2018). We provide an attack cost analysis and pre- and post-attack

regret analysis of Pareto UCB for MO-MAB herein.

The rest of the paper is as follows. In Section 2, we present the existing work in the domain of MO-MAB. Then in Section 3, we introduce the preliminaries including notations and formulations in the context of MO-MAB. We proceed by proposing an algorithm with optimality in both stochastic and adversarial MO-MAB, as well as a nearly optimal one up to a $\log T$ factor, and establishing their regret upper bounds in Section 4. Finally, in Section 5, we provide full analyses on the regret lower bound and on a special case when Pareto UCB is adopted.

2. Literature Review

MAB is well understood and many algorithms have proven to be effective for MAB both theoretically and numerically. For stochastic MAB, UCB and Thompson Sampling achieve optimal regret of order $\log T$. For adversarial MAB, (Auer et al., 2002) propose EXP3 with an optimal regret $\sqrt{T}$. Moreover, in the work of (Seldin and Slivkins, 2014), EXP3++ algorithm is designed as a nearly optimal solution to both stochastic and adversarial setting, up to a $\log T$ factor. More specifically, the EXP3++ algorithm dynamically updates two levers used for EXP3 and UCB, respectively, by estimating the assumed nature of rewards, and yields a sample rule accordingly that is a weighted average of EXP3 and UCB. The SAO algorithm (Auer and Chiang, 2016) improves it by having smaller regret in the stochastic setting but with a larger regret in the adversarial setting instead and with more complicated steps. (Seldin and Lugosi, 2017) improve on EXP3++ by a careful parameterization that leads to smaller regret and does not require a prior T and (Zimmert and Seldin, 2019) finally arrive at optimum for both settings under certainly strong conditions. MO-MAB extends the single-dimensional objectives of MAB to be multi-dimensional ones and makes it strikingly challenging to develop optimal methods. To this end, it is natural to extend MAB algorithms to MO-MAB and Pareto UCB is such a success. Specifically, Pareto UCB (Drugan and Nowe, 2013) constructs upper bounds of confidence intervals for each arm and determines a Pareto optimal front instead of the best arm in the presence of Pareto optimality. An arm is randomly pulled from the set. Pareto regret of order $\log T$ for Pareto UCB is guaranteed by the multi-dimensional Chernoff-Hoeffding bound. Further improvement on it can be found in (Drugan et al., 2014). However, they only work for stochastic MO-MAB. There are variants of stochastic MO-MAB, such as PF-LEX for MO-MAB with a lexicographical order and satisficing objectives (which is a relaxation of maximizing that allows for suboptimal less risky decision making in the face of uncertainty) (Hüyük and Tekin, 2021) and MO-linUCB for contextual MO-MAB (Mehrotra et al., 2020). But no attention has been given to the framework of adversarial MO-MAB.

Another line of work on MO-MAB is to modify Multi-objective Optimization (MOO) methods. A lot of algorithms have shown empirically great performance in real applications, such as larg-scale MOEA using evolutionary algorithms dealing with the classic trade-off between exploration and exploitation (Hong et al., 2021) and Annealing Pareto Knowledge Gradient as a variant of Thompson Sampling (Yahyaa et al., 2014). But the lack of a theoretical guarantee presents a concern since data-driven methods may be less robust and more sensitive, which motivates focus on theoretically effective algorithms. MOO methods consider both deterministic settings and settings under uncertainty, where the former can be adapted to stochastic MO-MAB with deterministic mean vectors and the latter raises the possibility of use in adversarial MO-MAB. For stochastic MO-MAB, various scalarization techniques from MOO are proposed for dimension reduction, such as Linear and Chebyshev, with which the problem essentially degenerates to MAB. But scalarizations may result in a loss of information and be highly problem-dependent as stated earlier. Optimizing the General Gini Index (GGI) using online convex optimization (Busa-Fekete et al., 2017) extends the concept of the Gini Index by means of ordered weighted average, albeit the regret metric is again defined on GGI. For MOO under uncertainty, utility functions including both linear and non-linear options are often applied in economics, whereas they work similarly as scalarizarion methods. Recently, (Groetzner and Werner, 2022) suggest a relative minimax regret approach with theoretical guarantees. Specifically, the minimax regret first takes maximum over the random variables for each dimension and then minimum over the decision variables using Pareto order relationships. However, it requires precomputing all optimal values with respect to random variables, which does not generalize to on-the-fly bandit settings. Herein we fill the gap by formally introducing the Pareto regret for adversarial MO-MAB and by proposing algorithms that show optimality or near optimality in both stochastic and adversarial MO-MAB.

Lower bound analyses on regret play an important role in studying the optimality of algorithms in MO-MAB. On the one hand, the work on MO-MAB algorithms usually shows lower bounds on regret depending on the problem settings and what regret metrics are used. For Pareto regret, (Drugan and Nowe, 2013) argue a lower bound of order $\log T$ for stochastic MO-MAB that matches both the regret upper bound of Pareto UCB and the regret lower bound for stochastic MAB. For our newly proposed Pareto regrets, we also establish their lower bounds in both stochastic and adversarial MO-MAB, which are consistent not only with most of the upper bounds, but with (Drugan and Nowe, 2013) in stochastic settings, which further supports the validity of the proposed regret metrics. On the flip side, an adversar-

ial attack on UCB (Jun et al., 2018) shows that UCB may suffer a linear regret given a small attack cost, while the vulnerability of Pareto UCB is not yet studied which potentially provides bad cases where Pareto UCB does not work. In this paper, we consider the adversarial attack on Pareto UCB, which not only justifies its use in MO-MAB, but as an evidence of the limitation of Pareto UCB for adversarial MO-MAB.

3. Preliminaries

In this section, we start with the notations used throughout the paper and then formally introduce Pareto regret including the existing one for stochastic MO-MAB and our newly proposed ones that capture both stochastic MO-MAB and adversarial MO-MAB. The regret analyses are presented in the next section.

3.1. Notation

For vectors, a Pareto order relationship has been introduced in (Drugan and Nowe, 2013) and (Drugan et al., 2014). There are several possible order relationships between two vectors a and b. Vector a is said to weakly dominate b, written as $a \succeq b$ or $b \preceq a$, if and only if for any dimension d, $a_d \geq b_d$. Removing the equality for at least one dimension gives us dominating, which is denoted by $a \succ b$ or $b \prec a$. Vector a is incomparable with b, i.e. $a||b$, if there exists a dimension d such that $a_d > b_d$ and another dimension d' satisfying $a_{d'} < b_{d'}$. We say that vector a is non-dominated by b if there is dimension d such that $a_d > b_d$.

We consider MO-MAB with K arms being $\{1, 2, \ldots, K\}$ and D-dimensional rewards and time horizon T. The chosen arm by a player at each time step t is denoted by a_t and only the reward of a_t, namely $r^{a_t,t}$, is revealed to the player. Value $N_i(t)$ is the total number of pulls of arm i by the player up to time t. In the stochastic setting, at each time step t, the reward $r^{i,t} = (r_l^{i,t})_{1 \leq l \leq D}$ of arm i follows a distribution with time-invariant mean vector $\mu^i = (\mu_l^i)_{1 \leq l \leq D}$ satisfying $1 \succeq \mu^i \succeq 0$. For adversarial MO-MAB, the reward $0 \preceq r^{i,t} \preceq 1$ of each arm i is specified by an adversary which can be adaptive or oblivious where the former depends on past actions of players and the latter does not.

3.2. Pareto Optimality and Regret

3.2.1. Optimality

The Pareto order relationship can determine the Pareto optimality of arms as follows. Formally, in stochastic settings, an arm j is said to be Pareto optimal, if μ^j is non-dominated by the reward mean vectors of any other arm and the set of Pareto optimal arms is named as the Pareto optimal set, denoted by O_A (Drugan and Nowe, 2013). We propose to define the Pareto optimal set O'_A and $\bar{O}'_A$ based on Pareto optimality with respect to the cumulative reward vector $\sum_{t=1}^T r^{i,t}$ and $\sum_{t=1}^T E[r^{i,t}]$, respectively, for both stochastic settings and adversarial settings. Formally, $O'_A = \{i^* : \sum_{t=1}^T r^{i^*,t}$ is non-dominated by $\sum_{t=1}^T r^{j,t}$, for any arm $j \neq i^*\}, \bar{O}'_A = \{i^* : \sum_{t=1}^T E[r^{i^*,t}]$ is non-dominated by $\sum_{t=1}^T E[r^{j,t}]$ for any arm $j \neq i^*\}$ allow us to define a uniformly effective Pareto regret for MO-MAB. Note that $\bar{O}'_A$ is actually the same as O_A in stochastic settings. We denote the corresponding Pareto optimal front as $O = \{\mu^a : a \in O_A\}$, $O' = \{\sum_{t=1}^T r^{a,t} : a \in O'_A\}$, and $\bar{O}' = \{\sum_{t=1}^T \mathbb{E}[r^{a,t}] : a \in \bar{O}'_A\}$.

3.2.2. Regret

We now proceed to introduce Pareto regrets by measuring the distance between the obtained rewards and the rewards of arms in the Pareto optimal set with a metric, consistent with in (Drugan and Nowe, 2013). In (Drugan and Nowe, 2013), Pareto regret for Stochastic MO-MAB is denoted by $R_T = \sum_{t=1}^T Dist(\mu^{a_t}, O) = \sum_{i=1}^K N_i(T) \cdot Dist(\mu^i, O)$, where the distance measure $Dist(a, O)$ between a vector $a \preceq \sigma$ for every $\sigma \in O$ and a set O is defined by $Dist(a, O) = \min_{\epsilon \geq 0}\{\epsilon : a + \epsilon 1 || \sigma$ for every $\sigma \in O\}$. Here $1 \in R^D$ is a vector of all 1.

Nevertheless, this definition of Pareto regret is with respect to O which requires the reward mean vectors to be constant over time that only holds for stochastic MO-MAB. To this end, we propose a Pareto regret R'_T and a Pareto pseudo regret $\bar{R}'_T$ based on a similar distance metric but with the newly defined Pareto optimal fronts. To our best knowledge, this is the first work on defining and establishing Pareto regret in adversarial MO-MAB with the generalizability to stochastic MO-MAB as well.

Pareto Regret Formally, the Pareto regret in MO-MAB is defined as $R'_T = Dist(\sum_{t=1}^T r^{a_t,t}, O')$.

Pareto Pseudo Regret In the context of MO-MAB, we propose a Pareto Pseudo regret as $\bar{R}'_T = Dist(E[\sum_{t=1}^T r^{a_t,t}], \bar{O}')$.

In the proofs and some statements we utilize rewards only along a single dimension. To this end, denote R_T^d as the regret with respect to rewards of dimension or coordinate d as $R_T^d = \max_i \sum_{t=1}^T r_d^{i,t} - \sum_{t=1}^T r_d^{a_t,t}$, and the corresponding pseudo regret is defined as $\bar{R}_T^d = \max_i E[\sum_{t=1}^T r_d^{i,t}] - E[\sum_{t=1}^T r_d^{a_t,t}]$.

Note that the proposed regrets apply for both stochastic MO-MAB and adversarial MO-MAB since rewards can be arbitrary. To this end, we call R'_T the general Pareto regret, different from R_T, namely the stochastic Pareto regret that only works for the stochastic setting.

4. Upper bounds on Pareto regret

In this section, we formally elaborate on the newly proposed algorithms for MO-MAB based on MAB algorithms, which

we consider both with and without a priori knowledge of stochastic or adversarial, and prove their optimality or near optimality in stochastic and adversarial settings simultaneously. More specifically, we denote an indicator for the setting of MO-MAB by s with $s = 0$ being stochastic and $s = 1$ being adversarial. When s is known, we propose Algorithm 1 that achieves optimality. For unknown s, i.e. with less information, Algorithm 2 is nearly optimal with respect to Pareto pseudo regret, up to a $\log T$ factor, in line with the work on MAB.

4.1. A Priori Knowledge of Stochastic or Adversarial

When the indicator s is given, Pareto regret in MO-MAB can be related to regret in MAB as follows.

Theorem 4.1. *For any dimension d', we have that $R'_T \leq R^{d'}_T$.*

Based on the result which essentially fully characterizes Pareto regret in the form of vanilla regret in MAB, we formally develop Algorithm 1 called Multi-Objective with Known S (MO-KS) for MO-MAB, which has comparable performance with MAB by switching between EXP3.P and UCB depending on the problem setting s and a randomly sampled dimension d'. Theorem 4.2 shows the Pareto regret upper bound of Algorithm 1 with respect to Pareto regret R'_T, which has the same order as the regret upper bound in MAB.

Algorithm 1 Algorithm With Known s (MO-KS)

Input: Fixed dimension d', $1 \leq d' \leq D$;
Initialization: indicator $s = \{0, 1\}$;
if s= 0 **then**
  for $t = 1, 2, \ldots, T$ **do**
    Play the bandit game by applying the UCB algorithm along dimension d'
  end for
else if s= 1 **then**
  for $t = 1, 2, \ldots, T$ **do**
    Play the bandit game by applying EXP3.P algorithm along dimension d'
  end for
end if

Theorem 4.2. *Based on Algorithm 1, for dimension d' we have $E[R^{d'}_T] \leq O^*(\log T) \cdot I_{s=0} + O^*(\sqrt{T}) \cdot I_{s=1}$, which leads to $E[R'_T] \leq O^*(\log T) \cdot I_{s=0} + O^*(\sqrt{T}) \cdot I_{s=1}$.*

The proof of Theorem 4.2 is by combining the result of Theorem 4.1 and the existing regret upper bounds of the UCB and EXP3.P algorithms.

Note that the choice of d' in Algorithm 1 is arbitrary and thus it can depend on the context, since the theoretical guarantee holds for any d'. For example, for a recommendation system, the click rate may be of more interest for decision makers and thereby being the optimization objective. Moreover, from the perspective of minimizing the constant term in the Pareto regret, we can always specify such a dimension accordingly by running a burning period to determine the optimal dimension, despite of the fact that the regret order remains the same which is a focus of this paper.

4.2. Lack of Knowledge of Stochastic or Adversarial

In this section, we focus on MO-MAB where the indicator s is unknown and propose an algorithm (see Algorithm 2) that remains nearly optimal, up to a $\log T$ factor, no matter what is the value of s. Moreover, Algorithm 2 allows unknown T, compared to the scenarios with known indicators, which increases generalizability.

Similarly to Theorem 1 we have the following result that makes it a possibility to analyze Pareto pseudo regret by means of pseudo regret in the context of MO-MAB.

Theorem 4.3. *For any dimension d', we have that $\bar{R}'_T \leq \bar{R}^{d'}_T$.*

Algorithm 2 Algorithm With Unknown s and T (MO-US)

Input: Fixed dimension d', $1 \leq d' \leq D$;
Initialization: $c = 256, \alpha = 3$;
Pull each arm once and set $\tilde{L}_a(K) = r^a_{d'}$ and $N_a(K) = 1$ for any arm a;
for $t = K+1, K+2, \ldots, T$ **do**
  $\eta_t = \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$;
  For any arm a let
  $UCB_a(t) = \min\{1, \frac{\tilde{L}_a(t-1)}{N_a(t-1)} + \sqrt{\frac{\alpha \ln tK^{\frac{1}{\alpha}}}{2N_a(t-1)}}\}$;
  $LCB_a(t) = \min\{1, \frac{\tilde{L}_a(t-1)}{N_a(t-1)} - \sqrt{\frac{\alpha \ln tK^{\frac{1}{\alpha}}}{2N_a(t-1)}}\}$;
  $\zeta_t(a) = \min\{0, LCB_a(t) - \min_{a'} UCB_{a'}(t)\}$;
  $\psi_t(a) = \frac{c \ln t}{t\zeta_t(a)^2}$; $\epsilon_t(a) = \min\{\frac{1}{2K}, \eta_t, \psi_t(a)\}$;
  $\rho_t(a) = \frac{\exp -(\eta_t \tilde{L}_{t-1}(a))}{\sum_{a'} \exp -(\eta_t \tilde{L}_{t-1}(a'))}$;
  $\tilde{\rho}_t(a) = (1 - \sum_{a'} \epsilon_t(a'))\rho_t(a) + \epsilon_t(a)$;
  Pull an arm a_t based on probability $\tilde{\rho}_t(a)$ and receive the reward $r^{a_t,t}$;
  For any arm a let
  $\tilde{l}^a_t = \frac{1 - r^{a_t,t}_{d'}}{\tilde{\rho}_t(a)} \cdot 1_{a=a_t}$; $\tilde{L}_t(a) = \tilde{L}_{t-1}(a) + \tilde{l}^a_t$;
  $N_a(t) = N_a(t-1) + 1_{a=a_t}$;
end for

The proof of Theorem 4.3 follows the same steps as the proof of Theorem 4.1 and substitutes the analysis on Pareto regret with the one on Pareto pseudo regret.

Motivated by the relationship between Pareto pseudo regret and pseudo regret as in Theorem 4.3, we develop Algorithm 2 (MO-US) by modifying the EXP3++ algorithm in

MAB as in (Seldin and Slivkins, 2014; Seldin and Lugosi, 2017). We are given a dimension and apply the parameter-specific EXP3++ algorithm to guarantee near optimality for pseudo regret in that dimension and consequently for Pareto pseudo regret.

The upper bound on Pareto pseudo regret of Algorithm 2 is formally stated as follows.

Theorem 4.4. *In Algorithm 2, without knowing the time horizon T, for dimension d' we have $\bar{R}_T^{d'} \leq O^*((\log T)^2) \cdot I_{s=0} + O^*(\sqrt{T}) \cdot I_{s=1}$. Pareto Pseudo regret $\bar{R}'_T$ can be bounded as $\bar{R}'_T \leq O^*((\log T)^2) \cdot I_{s=0} + O^*(\sqrt{T}) \cdot I_{s=1}$.*

In like manner, the proof utilizes the result of Theorem 4.3 and the known results of IEXP3++ algorithm as in (Seldin and Lugosi, 2017).

Similarly, the discussion aforementioned in the previous subsection on the choice of dimension applies here.

Remark 4.5. Note that there is an algorithm proposed by (Zimmert and Seldin, 2019) to improve the regret order $(\log T)^2$ in stochastic MAB, which has order $\log T$, exactly the same order as the lower bound. However, the additional assumptions of the algorithm may limit its application in MO-MAB, such as the condition of a unique best arm. For MO-MAB, it is possible that multiple best arms exist in one dimension.

For stochastic MO-MAB, (Drugan and Nowe, 2013; Drugan et al., 2014) provide a formulation of regret and subsequently propose optimal algorithms, however their algorithms and analyses do not hold for adversarial settings. While techniques in multi-objective optimization ((Hong et al., 2021; Yahyaa et al., 2014; Busa-Fekete et al., 2017; Groetzner and Werner, 2022)) adapt to any reward distributions, they either convert the MO-MAB problem to single objective MAB by scalarization and minimax or only work empirically. Our MO-MAB results establish algorithms and regret bounds for the adversarial setting, Figure 1.

5. Lower bounds on Pareto Regret

5.1. Lower Bounds

In this section, we show that for stochastic MO-MAB and adversarial MO-MAB, there exist scenarios where the regret of any randomized algorithms is exceeding certain lower bounds, with respect to both Pareto regret and Pareto pseudo regret. It validates the optimality of Algorithm 1 and the near optimality of Algorithm 2, up to constant factors. Moreover, the lower bounds stay the same with those for MAB, which further connects MO-MAB with MAB, in conjunction with the upper bound results.

Formally, the lower bounds on Pareto regret and Pareto pseudo regret in both stochastic MO-MAB and adversarial MO-MAB are summarized as follows.

Theorem 5.1. *For stochastic MO-MAB, there exists scenarios where the Pareto pseudo regret of any randomized algorithms is larger than $\log T$.*

Theorem 5.2. *For stochastic MO-MAB and small T, there exist scenarios where the expected Pareto regret of any randomized algorithms is larger than $\log T$.*

Theorem 5.3. *For adversarial MO-MAB, there exist scenarios where the Pareto pseudo regret of any randomized algorithms is larger than $\sqrt{T}$.*

Theorem 5.4. *Consider adversarial MO-MAB with repeating values for different dimensions in reward vectors, i.e. $r^{i,t} = (r_1^{i,t}, \ldots, r_1^{i,t})$. Then there exists scenarios where the Pareto regret of any randomized algorithms is larger than $\sqrt{T}$ with high probability.*

In the proofs of Theorems 5.3 and 5.4, we find instances where the regret in MO-MAB is equivalent to the regret in MAB in one dimension, namely the marginal regret. This can be done by letting either the reward vectors or the reward mean vectors have the same numbers for all dimensions. Then the lower bounds on marginal regret essentially hold for MO-MAB when the reward numbers in vectors meet the corresponding conditions in MAB. This argument does not hold for Theorems 5.1 and 5.2 by noting that the reward vectors do not have the same numbers for all dimensions out of stochasticity. This brings additional difficulties and thereby necessitates a new analytical approach. We herein utilize the multi-dimensional concentration inequalities to deal with stochasticity.

For stochastic MO-MAB, (Drugan and Nowe, 2013) establish a lower bound of order $\log T$ and our results are consistent with it, with respect to the newly defined Pareto regret and Pareto pseudo regret. This justifies the introduction of the proposed measures. Furthermore, for adversarial MO-MAB, our regrets have lower bounds of order $\sqrt{T}$, which are the same as in single dimensional MAB (Gerchinovitz and Lattimore, 2016; Bubeck et al., 2012). Meanwhile, the upper bounds of MO-KS have the same order as the lower bounds, implying optimality of MO-KS, while MO-US is nearly optimal up to a $\log T$ factor since the upper bound in the stochastic setting is $(\log T)^2$.

5.2. Lower bounds of Pareto UCB

We have already shown the lower bounds for general algorithms. Motivated by the online adversarial attack on UCB with small attack cost in an MAB setting as studied in (Jun et al., 2018), for Pareto UCB designed for stochastic MO-MAB, we go further and show that it can lead to linear regrets in terms of Pareto regret R_T and R'_T given adaptive adversaries by borrowing the adversarial attack mechanism in MAB. This analysis is made possible by our newly proposed framework of adversarial MO-MAB, highlighting the inef-

fectiveness of Pareto UCB in such scenarios. Specifically, we consider scenarios where the reward vectors of all arms are sub-Gaussian distributed with the same variance denoted by σ^2 and μ^K is dominated by any other arm and thus for every $i < K, 0 < \Delta_i = \max_d\{\mu_d^i - \mu_d^K\} = ||\mu^i - \mu^K||_\infty$. Generally speaking, the adversary, namely Alice, aims to manipulate Bob, the player, into pulling arm K very often while making small attacks under assumptions that 1) Bob does not know the presence of Alice, 2) the number of arms K and the time horizon T are known to Alice and Bob, 3) Alice knows that arm K is dominated by all other arms, and 4) Alice knows the exact algorithm Bob is using. Note that Alice does not necessarily know the specific arm a_t pulled by Bob at time step t, which generalizes the adversarial attack in MAB as in (Jun et al., 2018) that assume a known a_t.

Formally, value α_t is the cost of Alice to attack Bob at time step t. Parameter Δ_0 is determined by Alice and $\delta \in (0, 1)$. (We have already defined $\Delta_i, i \geq 1$ but not Δ_0.) Let $\hat{\mu}^i(t)$ be an estimator for μ^i up to time step t, while $\hat{\tilde{\mu}}^i(t)$, $\tilde{\mu}^i = \mu^i - \bar{\alpha}_t = \mu^i - \frac{\sum_{t=1}^T \alpha_t}{T}$ are the estimator and true value of the post-attack reward of arm i, respectively. The attack

Algorithm 3 Attack UCB

Initialization: $\beta(n) = \sqrt{\frac{2\sigma^2}{n} \log \frac{\pi^2 K n^2}{3\delta}}$;
for $t = 1, 2, \ldots, T$ **do**
  Bob chooses a_t according to the UCB algorithm;
  Simultaneously Alice computes a_t based on past $t-1$ observations and the UCB algorithm;
  The environment generates rewards for each arm i by sampling from a stochastic generator;
  Alice learns the pre-attack reward $r^{a_t,t}$ from the environment;
  By using $r^{a_t,t}$ Alice updates all $\hat{\mu}, \hat{\tilde{\mu}}$;
  $\alpha_t = 1_{a_t=K} \cdot \max\{0, \hat{\mu}^K(t) - 2\beta(N_K(t))$
    $-\Delta_0 - \hat{\tilde{\mu}}^{a_t}(t)\}$;
  Bob receives reward $r^{a_t,t} - \alpha_t$;
end for

algorithm in MAB is shown in Algorithm 3. At each time step, Alice predicts the choice a_t of Bob given by the UCB algorithm, and attacks $a_t \neq K$ to mislead Bob.

When it comes to Pareto UCB in the context of MO-MAB, randomness in the Pareto optimal front presents a concern since a_t is unpredictable. As a generalization, we propose Algorithm 4 that attacks Pareto UCB with small costs. Let $O', \bar{O}'_t$ be the post-attack Pareto optimal front, let O again be the pre-attack Pareto optimal front on μ^i and O_t be the estimator for O up to time step t. As before we similarly denote the quantities with subscript A to capture arm indices. More specifically, we modify the estimation for the best arm in UCB by estimating Pareto optimal front O^t and attack any arm in O_A^t if $K \notin O_A^t$ and determine the attack cost as a corruption by the Pareto order relationship.

5.2.1. Results on Stochastic Pareto Regret

Under Algorithm 4, we present a lower bound on stochastic Pareto Regret R_T defined in (Drugan and Nowe, 2013). We assume that Alice does not attack in the first $2K$ rounds, i.e. $\sum_{s=1}^{2K} \alpha_s = 0$. The result generalizes the adversarial attack on UCB to adversarial attack on Pareto-UCB, i.e. from MAB to MO-MAB. We first state Theorem 5.5 which provides an upper bound on the number of times Bob does not pull arm K and on the total attack cost.

Algorithm 4 Attack Pareto UCB

Initialization: $\beta(n) = \sqrt{\frac{2\sigma^2}{n} \log \frac{\pi^2 K n^2}{3\delta}}$;
for $t = 1, 2, \ldots, T$ **do**
  Bob computes the Pareto front O_t;
  The environment generates reward vector for each arm by sampling from a generator;
  Simultaneously Alice computes the Pareto front O_t based on past $t-1$ observations and the Pareto UCB algorithm;
  Alice learns the pre-attack reward $r^{i,t}$ for $i \in O_A^t$;
  By using $r^{i,t}, i \in O_A^t$, Alice updates $\hat{\mu}, \hat{\tilde{\mu}}$;
  if $K \in O_A^t$ **then**
    $\alpha_t = 0$
  else
    for $j \in O_A^t$ **do**
      $\bar{z}^j = \hat{\mu}^K - (2\beta(N_K(t)) + \Delta_0) \cdot 1$;
      $\hat{z}^j = \frac{N_j(t-1)(\hat{\mu}^j(t-1)) - \sum_{s=1}^{t-1} \alpha_s \cdot 1 + r^{j,t}}{N_j(t)}$;
    end for
  end if
  $\alpha_t = \max\{\max_{j \in O_A^t, d}\{N_j(t)(\hat{z}_d^j - \bar{z}_d^j)\}, 0\}$;
  Bob randomly samples arm a_t from O_A^t and receives reward $r^{a_t,t} - \alpha_t \cdot 1$;
end for

Theorem 5.5. *Let $K \geq \frac{3e^2\delta}{\pi^2}$. With probability $1 - D\delta$, for any $T > 2K$, under Algorithm 4, any non-target arm is pulled $O^*(\log T)$ times and the total attack cost is $(K-1)\left(2 + \frac{9\sigma^2}{\Delta_0^2} \log T\right) \cdot \max_i(\Delta_i + \Delta_0) + O^*(\log T)$.*

As a consequence, the Pareto regret R_T can be bounded by its definition, which is stated as Theorem 5.6.

Theorem 5.6. *With probability $1 - D\delta$, for any $T > 2K$ and $K \geq \frac{3e^2\delta}{\pi^2}$, the Pareto regret R_T of the Pareto UCB algorithm is at least of order $O^*(T)$ under the adversarial attack generated by Algorithm 4.*

The proofs of Theorem 5.5 and Theorem 5.6 follow the logic of what have been established in (Jun et al., 2018), but

the analyses are mostly in the reward vector space and under our newly proposed attack algorithm. The dimensionality of rewards presents non-trivial challenges in that the attack is on multiple arms at each time step and Bob's rule of playing is Pareto UCB without deterministic choices. To this end, we use multi-dimensional concentration inequalities, (Drugan and Nowe, 2013).

For stochastic MO-MAB under no adversarial attack, Pareto UCB is optimal with respect to the Pareto regret R_T as established in (Drugan and Nowe, 2013), the optimality of which is unknown in adversarial settings. While in (Jun et al., 2018), the adversarial attack on UCB in MAB is fully studied, it does not apply to Pareto UCB for MO-MAB. When adversarial attack is added to Pareto UCB, which leads to an adversarial MO-MAB setting, we show that its Pareto regret R_T is at least linear in T.

5.2.2. Results on General Pareto Regret

Next, we study the lower bound on the general Pareto regret R'_T for Pareto UCB under Algorithm 4, as a validation for the claim that Pareto UCB is not optimal for adversarial MO-MAB.

Let us denote quantity $\bar{\alpha}_t^j$ to be the counter-factual attack cost of arm j with respect to arm K. They are computed by Alice only for the arms in the Pareto optimal set and are 0 for the arms not in the Pareto optimal set. In other words, the counter-factual attack cost $\bar{\alpha}_t^j$ is defined for each arm j as $\bar{\alpha}_t^i = \max\{\max_d N_i(t)(\hat{z}_d^i - \bar{z}_d^i), 0\}$. Based on Algorithm 4, $\alpha_t = \max_{j \in O_A^t} \bar{\alpha}_t^j = \max_j \bar{\alpha}_t^j$.

We start by properly defining the Pareto optimal front for stochastic MO-MAB with the attack cost, which is necessary for Pareto regret R'_T. Two options can be considered: one averaging the actual attack costs and the second taking into account the counterfactual attack costs.

Definition 1 Given adversarial attack cost, $\bar{O}'$ is defined as the Pareto optimal front of vectors $\mu^i - \frac{1}{\sum_{j \neq K} N_j(T)} \cdot \sum_{a_t \neq K} \alpha_t \cdot 1$ over all arm i, whereas O' is over vectors $\frac{1}{T}\sum_{t=1}^T r^{i,t} - \frac{1}{N_i(T)} \sum_{i_t = i} \alpha_t \cdot 1$.

Definition 2 Given adversarial attack cost, $\bar{O}'$ is defined as the Pareto front of vectors $\mu^i - \frac{1}{T}\sum_{t=1}^T \bar{\alpha}_t^i \cdot 1$, whereas O' is over $\frac{1}{T}(\sum_{t=1}^T r^{i,t} - \sum_{t=1}^T \bar{\alpha}_t^i \cdot 1)$.

Assumption 4 Let S be the set of $\{\mu^1, \ldots, \mu^K\}$. We assume that the distance from the arms that are not in Pareto front of S to the arms in Pareto front of S is at least 5γ for $0 < \gamma < \frac{1}{5}$. Formally, $\mu^j \succeq \mu^i + 5\gamma \cdot 1$ holds for any arm i not in Pareto front of S and arm j in Pareto front of S.

Based on either definition, we establish a lower bound on Pareto regret under certain conditions as follows. First, we consider the high probability regret lower bound which essentially implies a lower bound on the expected regret by integration.

Theorem 5.7. *Under Definition 1 and Assumption 4, with probability* $1 - 2\eta - D\delta$, *for any* T, $0 < \mu < 1$, $\delta \leq \frac{2\pi^2}{3e^2}$, *and* $\sigma^2 \leq \frac{1}{5(\ln(4D) + \ln\frac{1}{\mu})}$, *the Pareto regret* R'_T *of the Pareto UCB algorithm is at least of order* $O(T)$ *based on Algorithm 4, if* $K = 2$ *and* $\gamma \leq \sqrt{2\sigma^2(\ln(4D) + \ln\frac{1}{\mu})}$.

The proof connects R'_T and R_T by concentration inequalities and an asymptotic property of the attack cost and then studying R_T. Next based on Definition 2, we obtain a similar result.

Theorem 5.8. *Under Definition 2 and Assumption 4, with probability* $(1 - 2\eta - D\delta)$, *for* $T \geq T(\gamma)$, $0 < \mu < 1$, $\delta \leq \frac{2\pi^2}{3e^2}$ *and* $\sigma^2 \leq \frac{1}{5(\ln(4D) + \ln\frac{1}{\mu})}$, *the Pareto regret* R'_T *of the Pareto UCB algorithm is at least of order* $O(T)$ *based on Algorithm 4 if* $K = 2$ *and* $\gamma \leq \sqrt{2\sigma^2(\ln(4D) + \ln\frac{1}{\mu})}$.

Again, for the proof we utilize concentration inequalities and the asymptotic property of attack cost and the lower bound on R_T aforementioned.

For Pareto regret in (Drugan and Nowe, 2013), we previously establish that its lower bound is linear when attacking Pareto UCB with adversarial costs by extending the MAB result in (Jun et al., 2018). The regret analysis (Jun et al., 2018) only covers pre-attack rewards, while the post-attack scenarios are not studied. Here we generalize the result to our newly defined Pareto regret and Pareto pseudo regret defined on post-attack reward vectors, to further validate non-optimality of Pareto UCB in adversarial MO-MAB.

Remark 5.9. It is worth highlighting the robustness of our newly proposed algorithm against this online adversarial attack. In the presence of such attacks, where the attack cost is of order $\log T$ as shown in Theorem 5.5 and supported by Lemma 8.5, we observe that for any large t, $N_i(t) \leq \min\{N_K(t), 2 + \frac{9\sigma^2}{\Delta_0^2}\log t\}$. This implies that $N_K(t) = O(T)$. As a result, for at least $O(T)$ rounds, Alice refrains from launching attacks. During this period, the regret of our proposed algorithms is of order $O(\sqrt{T})$, which is different from Pareto UCB that exhibits linear regret in such scenarios.

6. Conclusion

In this paper, we study multi-objective multi-armed bandit (MO-MAB) with full Pareto regret analyses. We formulate the general MO-MAB problem for both stochastic and adversarial settings from a perspective of Pareto optimality and regret. Then we propose an optimal algorithm when given a priori knowledge of stochastic and adversarial and develop a nearly optimal algorithm up to a $\log T$ factor when lack of such knowledge is present. The theoretical guarantee is fully examined by both upper bounds and lower bounds on Pareto regret and Pareto pseudo regret. Moreover, we

extend the adversarial attack algorithm for UCB to a new one attacking Pareto UCB in the context of MO-MAB. We also establish a lower bound on stochastic Pareto regret for Pareto UCB and subsequently show that Pareto UCB may have poor performance in adversarial settings by providing a lower bound on general Pareto regret. A summary of the research gaps we close is shown in Figure 1 where the highlighted parts are the contributions of this paper.

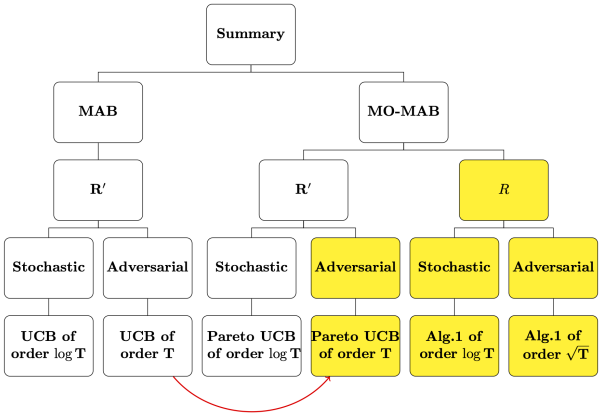

Figure 1: An illustration of our contributions

Table 1: Comparisons

Area	Application	Results
(1)	Pareto order relationship; Pareto optimal front; MO-MAB	$\sqrt{T \ln T}$
(2)	Attacks on stochastic bandits	regret T
(3)	(nearly) optimal in the (stochastic) adversarial settings	$\ln T$ ($\ln^2 T$); $\sqrt{T}$

Area	Limitation	Our results
(1)	scalarization functions; solving T subproblems	formulation of general MO-MAB
(2)	single-objective; regret not in adversarial settings	analyses in MO-MAB; regret T
(3)	single-objective	analyses in MO-MAB; $\ln T$ ($\ln^2 T$); $\sqrt{T}$

To explore and compare the connections between the research domains discussed herein, we present a comprehensive summary table (Table 1) that examines three distinct areas of research: Pareto optimization (labeled as (1)), adversarial attack (labeled as (2)), and best-of-both-worlds algorithms (labeled as (3)). We compare these methods from various perspectives, including their applications, known results, and areas where they have not yet been studied, or equivalently, potential limitations, as well as our own contributions to these fields. For simplicity, we use the term 'nearly optimal' to denote performance up to a factor of $\log T$.

As a future work, we point out that it could be of great interest to improve the upper bound of Pareto pseudo regret to achieve optimality given no knowledge of MO-MAB settings. Furthermore, this is in line with the current work on MAB, since it remains open to develop an optimal algorithm for general MAB under no assumptions. Meanwhile, it is worth investigating the explicit relationship between the regret upper bound and both the dimension D and the size of the Pareto optimal set $|O|$. The existing lower bound results already indicate a dependency on D, suggesting the possibility of enhancing the regret upper bound through the development of more effective algorithms. As a concluding remark, other metrics can be developed for measuring the performance of algorithms, such as unfairness.

7. Acknowledgement

We would like to thank the ICML anonymous reviewers and meta-reviewer for their helpful suggestions and valuable comments. Their feedback has greatly helped to improve the paper.

References

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, volume 7, pages 1–122, 2009.

P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120. Proceedings of Machine Learning Research, 2016.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

R. Busa-Fekete, B. Szörényi, P. Weng, and S. Mannor. Multi-objective bandits: Optimizing the generalized Gini index.

In *International Conference on Machine Learning*, pages 625–634. Proceedings of Machine Learning Research, 2017.

M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2013.

M. M. Drugan, A. Nowé, and B. Manderick. Pareto upper confidence bounds algorithms: an empirical study. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 1–8. IEEE, 2014.

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

P. Groetzner and R. Werner. Multiobjective optimization under uncertainty: A multiobjective robust (relative) regret approach. *European Journal of Operational Research*, 296(1):101–115, 2022.

W.-J. Hong, P. Yang, and K. Tang. Evolutionary computation for large-scale multi-objective optimization: A decade of progresses. *International Journal of Automation and Computing*, 18(2):155–169, 2021.

A. Hüyük and C. Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110(6):1233–1266, 2021.

K.-S. Jun, L. Li, Y. Ma, and J. Zhu. Adversarial attacks on stochastic bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

R. Mehrotra, N. Xue, and M. Lalmas. Bandit based optimization of multiple objectives on a music streaming platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3224–3233, 2020.

Y. Seldin and G. Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. Proceedings of Machine Learning Research, 2017.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. Proceedings of Machine Learning Research, 2014.

K. Van Moffaert, K. Van Vaerenbergh, P. Vrancx, and A. Nowé. Multi-objective χ-armed bandits. In *2014 International Joint Conference on Neural Networks*, pages 2331–2338. IEEE, 2014.

S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Annealing-pareto multi-objective multi-armed bandit algorithm. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 1–8. IEEE, 2014.

Y. Zhu and R. Nowak. On regret with multiple best arms. *Advances in Neural Information Processing Systems*, 33: 9050–9060, 2020.

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475. Proceedings of Machine Learning Research, 2019.

8. Appendix

8.1. Proofs of the results in Section 4

Throughout the following lemmas, we let a be a vector such that the set $\{\epsilon \geq 0 : a + \epsilon 1 || \sigma \text{ for every } \sigma \in O\}$ is not empty.

Lemma 8.1. *For any a and O we have*

$$u = Dist(a, O) = \min_d \max_{\sigma \in O}(\sigma_d - a_d).$$

Proof. By definition, we have that

$$\begin{aligned} Dist(a, O) &= \min_{\epsilon \geq 0}\{\epsilon : a + \epsilon 1 || \sigma \text{ for every } \sigma \in O\} \\ &= \min_{\epsilon \geq 0}(A_\epsilon). \end{aligned}$$

For $\epsilon \in A_\epsilon$ and for every $\sigma \in O$, there exists at least one dimension d such that $a_d + \epsilon > \sigma_d$ and one dimension d' such that $a_{d'} + \epsilon \leq \sigma_{d'}$.

Let B be the set of ϵ where for every $\sigma \in O$, there exists a dimension d such that $\epsilon \leq \sigma_d - a_d$. Similarly, let C be the set of ϵ where for every $\sigma \in O$, there exists a dimension d' such that $\epsilon > \sigma_{d'} - a_{d'}$. Clearly, we observe that

$$A_\epsilon = B \cap C.$$

Formally, we have $B = \{\epsilon : 0 \leq \epsilon \leq b = \min_\sigma \max_d(\sigma_d - a_d)\}$ and $C = \{\epsilon : \epsilon \geq c = \min_d \max_\sigma(\sigma_d - a_d)\}$, by their definitions. Note that for $A_\epsilon \neq \emptyset$, it must be the case that $c \leq b$. Therefore, it indicates

$$\min_\epsilon(A_\epsilon) = \min_\epsilon(B \cap C) = c,$$

which completes the proof.

□

Lemma 8.2. *If $O = \{\sigma\}$, then $Dist(a, O) = \min_d(\sigma_d - a_d)$.*

Proof. This is a consequence of letting $O = \{\sigma\}$ in Lemma 8.1.

□

Lemma 8.3. *There exists $\bar{\sigma} \in O$ such that $a + u1 \preceq \bar{\sigma}$ where u is defined in Lemma 8.1. For any dimension d we have*

$$u \leq \max_{\sigma \in O}(\sigma_d - a_d).$$

Proof. We prove the first part by contradiction. Let us assume that $a + u1 || \sigma$ for every $\sigma \in O$. Then for each σ there exist $d = d(\sigma)$ such that $a_d + u > \sigma_d$. We denote by $s(\sigma)$ the set of all dimensions with this property. We have $s(\sigma) \neq \emptyset$, $(s(\sigma))^c \neq \emptyset$. Let us define $l(\sigma) = \max_{d \in s(\sigma)}(a_d + u - \sigma_d)$ and $d^*(\sigma) = \arg\max_{d \in s(\sigma)}(a_d + u - \sigma_d)$. By definition $l(\sigma) > 0$. Let $\epsilon = \min_{\sigma \in O} l(\sigma) > 0$.

We consider $a + (u - \frac{\epsilon}{2})1$. For every $\sigma \in O$ and $d \in d^*(\sigma) \neq \emptyset$, we have $a_d + u - \frac{\epsilon}{2} \geq a_d + u - \frac{l(\sigma)}{2} = a_d + u - \frac{a_d + u - \sigma_d}{2} > \sigma_d$. For $d \notin s(\sigma)$ (note that $s(\sigma)^c \neq \emptyset$), we have $a_d + u - \frac{\epsilon}{2} \leq a_d + u \leq \sigma_d$.

We conclude that $a + (u - \frac{\epsilon}{2})1 || \sigma$ for every $\sigma \in O$ which contradicts the definition of u.

For the second part, $a + u1 \preceq \bar{\sigma}$ implies that for every d we have $u \leq \bar{\sigma}_d - a_d$. The existence of $\bar{\sigma}$ gives us that $u \leq \max_{\sigma \in O}(\sigma_d - a_d)$.

□

8.1.1. Proof of Theorem 4.1

Proof. It follows from Lemma 8.3 and definitions. Specifically, we use Lemma 8.3 by applying $a = \sum_{t=1}^T r^{a_t, t}$ and $O = \sum_{t=1}^T r^{i^*, t}$, where a meets the condition that $a \notin \{a \succeq \sigma \text{ for some } \sigma \in O\}$. Otherwise, by the fact that $\sum_{t=1}^T r^{a_t, t}$ does not dominate any $\sum_{t=1}^T r^{i^*, t} \in O$ which is the set of Pareto optimal reward vectors, we have $a + c1 = \sigma$ for exactly one c and $\sigma \in O$. In this case, the distance is equivalent to $\sigma_d - a_d \leq \max_{\sigma \in O}(\sigma_d - s_d)$.

□

8.1.2. Proof of Theorem 4.2

Proof. The result can be shown by combining the regret analysis of the EXP3.P and UCB algorithms in one-dimension MAB.

Formally, when $s = 0$, regret $R_T^{d'}$ of UCB satisfies that $E[R_T^{d'}] \leq O^*(\log T)$.

While $s = 1$ gives us that the regret of EXP3.P satisfies $E[R_T^{d'}] \leq O^*(\sqrt{T})$.

Meanwhile, according to Theorem 4.1, we have that $R_T \leq \min_d R_T^d$. Then by the Jensen's inequality, the expected value of R_T can be upper bounded by

$$E[R_T] \leq E[\min_d R_T^d] \leq \min_d E[R_T^d] \leq E[R_T^{d'}]. \quad (1)$$

To conclude, the result follows by plugging the upper bounds on $E[R_T^{d'}]$ in (1).

□

8.1.3. Proof of Theorem 4.3

Proof. It follows from Lemma 8.3 and definitions. Likewise, we again leverage Lemma 8.3 by using $a =$

$\sum_{t=1}^{T} E[r^{a_t,t}]$ and $O = \sum_{t=1}^{T} E[r^{i^*,t}]$, where a meets the condition that $a \notin \{a \succeq \sigma \text{ for some } \sigma \in O\}$. Otherwise, by noting that $\sum_{t=1}^{T} E[r^{a_t,t}]$ does not dominate any $\sum_{t=1}^{T} E[r^{i^*,t}] \in O$ which is the set of Pareto optimal expected reward vectors, we obtain $a + c1 = \sigma$ for exactly one c and $\sigma \in O$ and thus the distance equals to $\sigma_d - a_d \leq \max_{\sigma \in O}(\sigma_d - s_d)$, which completes the proof.

□

8.1.4. Proof of Theorem 4.4

Proof. By the choice of parameters $\eta_t, \zeta_t(a), \psi_t(a)$, the results hold as follows. In the adversarial regime, we have that the pseudo regret satisfies

$$\bar{R}_T^{d'} \leq O^*(\sqrt{T}),$$

while in the stochastic regime, the pseudo regret satisfies

$$\bar{R}_T^{d'} \leq O^*((\log T)^2).$$

Meanwhile, we have that

$$\bar{R}'_T \leq \min_d \bar{R}_T^d \leq \bar{R}_T^{d'}, \tag{2}$$

where the first inequality holds by the argument as in the proof of Theorem 4.3.

By plugging the upper bounds of $\bar{R}_T^{d'}$ in 2, we derive the statement in the theorem.

□

8.2. Proofs of the results in Section 5

8.2.1. Proof of Theorem 5.1

Proof. We consider stochastic MO-MAB with reward vectors $\mu^i = (\mu_1^i, \ldots, \mu_1^i) \in R^D$ for any arm i and any time steps t. By the definition of $\bar{O}'$, we have that $\bar{O}'$ is equivalent to a unique best arm i^* with $\mu_1^{i^*} = \max_i \mu_1^i$. It gives us that $\bar{R}'_T = \bar{R}_T^1 = \ldots = \bar{R}_T^D$ by noting that for any dimension d and by Lemma 2 we have

$$\begin{aligned}
\bar{R}'_T &= Dist(E[\sum_{t=1}^{T} r^{a_t,t}], \bar{O}') \\
&= Dist(E[\sum_{t=1}^{T} r^{a_t,t}], \{T\mu^{i^*}\}) \\
&= Dist(\sum_{t=1}^{T} \mu^{a_t}, \{T\mu^{i^*}\}) \\
&= T \cdot \mu_d^{i^*} - \sum_{t=1}^{T} \mu_1^{a_t} = \bar{R}_T^d.
\end{aligned}$$

By the result in (Bubeck et al., 2012) stating that

$$\liminf_T \frac{\bar{R}_T^1}{\log T} \geq O(1),$$

we have

$$\liminf_T \frac{\bar{R}'_T}{\log T} \geq O(1).$$

□

8.2.2. Proof of Theorem 5.2

Proof. We can assume $N_i(T) \geq 1$ since otherwise the underlying algorithm has linear regret. Consider stochastic MO-MAB with reward mean vectors being $\mu^i = \mu_1^i \cdot 1$ for any arm i and any time steps t. It indicates that $r_j^{i,t}, r_k^{i,t}$ are i.i.d for any $1 \leq j, k \leq D$, though potentially having different values as a result of stochasticity. Therefore, the difference between R'_T and R_T^1 can be bounded by the concentration inequality as follows.

Note that by Lemma 8.2 and 8.3, Theorem 4.1 and Theorem 4.3,we have

$$\begin{aligned}
R'_T &= Dist(\sum_{t=1}^{T} r^{a_t,t}, O') \\
&= \min_d \max_i (\sum_t r_d^{i,t} - \sum_{t=1}^{T} r_d^{a_t,t}) \\
&= \min_d \max_i (\sum_{t=1}^{T} r_d^{i,t} - \sum_{j=1}^{K} \sum_{a_t=j} r_d^{j,t})
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
\bar{R}'_T &= Dist(E[\sum_{t=1}^{T} r^{a_t,t}], \bar{O}') \\
&= \min_d \max_i E[\sum_t r_d^{i,t} - \sum_{t=1}^{T} r_d^{a_t,t}].
\end{aligned}$$

We use the Hoeffding's inequality, which leads to

$$\begin{aligned}
P(C_A) &= P(\forall i : |\mu_d^i - \frac{\sum_t r_d^{i,t}}{N_i(T)}| \leq |\frac{K^4 \log\log N_i(T)}{N_i(T)}|) \\
&\geq 1 - \eta
\end{aligned} \tag{4}$$

where $\eta = \sum_i 2\exp\{-2N_i(T) \cdot (\frac{K^4 \log\log N_i(T)}{N_i(T)})^2\} = \sum_i 2\exp\{-2 \cdot \frac{(K^4 \log\log N_i(T))^2}{N_i(T)}\}$.

Through the end of the proof, we assume that $K = 2$ and $11 < T \leq 1200$, i.e. 2-arm bandits with mean values μ^1, μ^2.

We first argue that $f(x) = \frac{(\log\log x)^2}{x}$ is decreasing for $x \geq 11$. We get $f'(x) = \frac{\log\log x}{x^2} \cdot (\frac{2-(\log\log x)\log x}{\log x})$. For $x \geq 11$, $\log x \geq 0$, $\log\log x \geq 0$ and both are increasing. We conclude that $(\log\log x)\cdot\log x$ is increasing for $x \geq 11$. Since $f'(11) < 0$, it then follows that $f'(x) < 0$ for $x \geq 11$.

If there exists arm i such that $N_i(T) = 1$, then we have $N_j(T) = T - 1$ for $j \neq i$ since $K = 2$.

In this case $\frac{(K^4 \log\log N_i(T))^2}{N_i(T)} = \infty$, which implies that

$$\begin{aligned}
\eta &= 2\exp\{-\infty\} + 2\exp\{-2\frac{(K^4\log\log(T-1))^2}{T-1}\} \\
&= 2\exp\{-2\frac{(K^4\log\log(T-1))^2}{T-1}\} \\
&\leq 2\exp\{-2\frac{(K^4\log\log T)^2}{T}\} \\
&= 2\exp\{-2^9\frac{(\log\log T)^2}{T}\} \\
&\leq 2\cdot\exp(-2^9\cdot\frac{(\log\log 1200)^2}{1200}) \leq 2\cdot\frac{1}{5} = \frac{2}{5}
\end{aligned}$$

where the first inequality is by monotonicity of $\frac{(\log\log x)^2}{x}$ for $x \geq 11$.

Suppose now that for $i = 1, 2$ we have $N_i(T) \geq 2$, and $N_i(T) \leq T$. For $x = 2, 3, \ldots, 1200$, $f(x)$ can only have maximum in $x \in \{2, 3, \ldots, 10, 1200\}$. Its maximum is for $x = 3$ and then the maximum in the η term is at $N_i(T) = 3$.

We thus have

$$\eta \leq 2\cdot 2\exp\{-2^9\frac{(\log\log 3)^2}{3}\} < \frac{8}{9}.$$

Therefore, we have $1 - \eta \geq \frac{1}{9} > 0$. With probability $1 - \eta$,

$$\begin{aligned}
R'_T &= \min_d \max_i (\sum_t r_d^{i,t} - \sum_{j=1}^K \sum_{a_t=j} r_d^{j,t}) \\
&\geq \min_d \max_i (T\cdot\mu_d^i - K^4\log\log T - \\
&\qquad \sum_{t=1}^T \mu_d^{a_t,t} - \sum_{j=1}^K |K^4\log\log Nj(T)|) \\
&\geq \min_d \max_i (T\cdot\mu_d^i - \sum_{t=1}^T \mu_d^{a_t,t} - (K+1)K^4\cdot\log\log T) \\
&= \min_d \bar{R}_T^d - (K+1)K^4\log\log T \\
&= \min_d \bar{R}_T^d - 48\log\log T
\end{aligned}$$

where the first inequality is straightforward from (4) and the second inequality holds by noticing $N_i(T) \leq T$.

Consider the instance-dependent lower bound with $\theta = \mu_1^2 - \mu_1^1 = \mu_2^2 - \mu_2^1 > 0$. By the result in (Lattimore and Szepesvári, 2020) for MAB, we have that for any T and any algorithm with $\sqrt{T}$ regret upper bound, it must hold

$$\bar{R}_T^d \geq \frac{\log T}{\theta}.$$

When choosing $\theta \leq \frac{1}{48}$, we derive

$$\begin{aligned}
R'_T &\geq \bar{R}_T^d - 48\log\log T \qquad (5)\\
&\geq 48(\log T - \log\log T). \qquad (6)
\end{aligned}$$

To conclude, we obtain

$$\begin{aligned}
E[R'_T] &\geq E[R'_T \cdot I_{C_A}] \\
&\geq 48(\log T - \log\log T)\cdot(1-\eta) \\
&\geq \frac{48}{9}O^*(\log T) = O^*(\log T)
\end{aligned}$$

where the second inequality holds by (6) and the last inequality uses $1 - \eta > \frac{1}{9}$.

□

8.2.3. Proof of Theorem 5.3

Proof. (Bubeck et al., 2012) analyze the lower bound on pseudo regret for MAB and establish that for any $\epsilon > 0$, there exists an adversarial MAB, denoted by $(r_1^{i,t})$, satisfying

$$\inf \bar{R}_T^1 \geq O^*(\sqrt{T}) - \epsilon \qquad (7)$$

where inf is taken over all algorithms.

Again, we focus on adversarial MO-MAB with reward vectors constant in dimensions, i.e. $r^{i,t} = (r_1^{i,t}, \ldots, r_1^{i,t})$, which is equivalent to MAB in the sense that $\bar{R}'_T = \bar{R}_T^1 = \ldots = \bar{R}_T^D$ given by

$$\bar{O}'_A = \{i^*\} = \arg\max_i E[\sum_{t=1}^T r_d^{i,t}]$$

and subsequently by Lemma 8.2, we have

$$\begin{aligned}
\bar{R}'_T &= Dist\left(E[\sum_{t=1}^T r^{a_t,t}], \bar{O}'\right) \\
&= Dist\left(E[\sum_{t=1}^T r^{a_t,t}], \{E[\sum_{t=1}^T r_d^{i^*,t}]\}\right) \\
&= E[\sum_{t=1}^T r_d^{i^*,t}] - E[\sum_{t=1}^T r_d^{a_t,t}] = \bar{R}_T^d
\end{aligned}$$

for any $1 \leq d \leq D$.

This is to say that $\inf \bar{R}'_T \geq O^*(\sqrt{T})$ by (7) holds for all algorithms.

□

8.2.4. Proof of Theorem 5.4

Proof. Consider adversarial MO-MAB with reward vectors being $r^{i,t} = (r_1^{i,t}, \ldots, r_1^{i,t})$. The Pareto optimal set O'_A is the unique best arm i^* that satisfies $i^* = \arg\max_i \sum_{t=1}^T r_1^{i,t}$. Then the MO-MAB problem essentially degenerates to MAB since $R'_T = R^1_T = \ldots = R^D_T$ which is guaranteed by

$$R'_T = Dist\left(\sum_{t=1}^T r^{a_t,t}, O'\right) = \sum_{t=1}^T r_d^{i^*,t} - \sum_{t=1}^T r_d^{a_t,t} = R^d_T, \text{ for every } d.$$

By the result in (Gerchinovitz and Lattimore, 2016) stating that

$$R^1_T \geq O^*(\sqrt{T})$$

holds for any randomized algorithm, we have that the Pareto regret R'_T is larger than $\sqrt{T}$.

□

8.2.5. Results in Section 5.2

8.2.6. Proof of Theorem 5.5

Proof of Theorem 5.5. With $\beta(n) = \sqrt{\frac{2\sigma^2}{n}\log\frac{\pi^2 K n^2}{3\delta}}$ where σ^2 is the variance of values for different dimensions in reward vectors of different arms, let us define event E as

$$E = \{\forall i, \forall t > K : ||\hat{\mu}^i(t) - \mu^i||_\infty < \beta(N_i(t))\}.$$

We first present several lemmas used later. As before, without loss of generality we assume that in the first K steps every arm is pulled. This implies $N_i(t) \geq 1$ for any $t > K$ and arm i.

Lemma 8.4. *For any $\delta \in (0,1)$, $P(E) > 1 - D\delta$.*

Proof. Note that

$$\begin{aligned}
&P(E)\\
&= P(\forall i, \forall t > K : ||\hat{\mu}^i(t) - \mu^i||_\infty < \beta(N_i(t)))\\
&= P(\forall i, \forall t > K, \forall d : \\
&\qquad |N_i^{-1}(t)\sum_{a_s=i} r_d^{i,s} - \mu_d^i| < \beta(N_i(t)))\\
&\geq 1 - 2\sum_{t>K}\sum_{i\leq K} D\exp\{-\frac{N_i(t)}{2\sigma^2}\beta(N_i(t))^2\}\\
&= 1 - 2\sum_{t>K}\sum_{i\leq K} D\exp\{-\frac{N_i(t)}{2\sigma^2}\frac{2\sigma^2}{N_i(t)}\log\frac{\pi^2 K N_i^2(t)}{3\delta}\}\\
&= 1 - 2\sum_{t>K}\sum_{i\leq K} D\left(\frac{\pi^2 K N_i^2(t)}{3\delta}\right)^{-1}\\
&= 1 - \frac{6\delta}{\pi^2 K} D\sum_{t>K}\sum_{i\leq K} N_i^{-2}(t)\\
&\geq 1 - D\delta\cdot\frac{6}{\pi^2 K}\frac{K\pi^2}{6}\\
&= 1 - D\delta
\end{aligned} \tag{8}$$

where the first inequality is by the Hoeffding's inequality for sub-Gaussian distributions and the last inequality holds by the fact that $\sum_{t=1}^\infty \frac{1}{t^{-2}} = \frac{\pi^2}{6}$.

□

Lemma 8.5. *Let us assume event E holds. Then, for any $i < K$ and any $t \geq 2K$, we have*

$$N_i(t) \leq \min\{N_K(t), 2 + \frac{9\sigma^2}{\Delta_0^2}\log t\}.$$

Proof. Since $t > 2K$, if $N_j(t) \leq 2$ for any $j < K$, then $N_K(t) \geq 2$ and the result holds trivially. Let $S = \{j < K : N_j(t) > 2\} \neq \emptyset$. If we prove the statement for each arm in S, this implies $N_K(t) > 2$ and thus the statement holds also for any arm $j \notin S$. Let now $i \in S$, i.e. $N_i(t) > 2$.

If $a_t \neq i$, then we can consider the last time t' we had $a_{t'} = i$ and apply the result in this case.

Let us assume $a_t = i$ and we consider the previous time step $t' < t$ when Bob pulled arm i. Since $a_t = i$, it implies $i \in O_A^t$. We have that $N_i(t'-1)+1 = N_i(t') = N_i(t)-1$, $N_i(t-1) = N_i(t')$ by the definition of $N_i(t)$. We note that by definition of $\alpha_{t'}$ we have $\bar{z}^j - \hat{z}^j + \frac{\alpha_{t'}}{N_j(t')} \geq 0$ for any $j \in O_A^{t'}$. This implies since $i \in O_A^{t'}$ and $\hat{z}^i - \frac{\alpha_{t'}}{N_i(t')} = \hat{\tilde{\mu}}^i(t')$, that

$$\hat{\tilde{\mu}}^i(t) \prec \hat{\tilde{\mu}}^K(t) - (2\beta(N_K(t)) + \Delta_0)\cdot 1. \tag{9}$$

Since $a_t = i$ is chosen by Bob based on Pareto UCB, there exists at least one dimension d, such that

$$\hat{\hat{\mu}}_d^i(t-1) + 3\sigma\sqrt{\frac{\log t}{N_i(t-1)}} \geq \hat{\hat{\mu}}_d^K(t-1) + 3\sigma\sqrt{\frac{\log t}{N_K(t-1)}}$$

and thus

$$\hat{\hat{\mu}}_d^i(t') + 3\sigma\sqrt{\frac{\log t}{N_i(t')}} \geq \hat{\hat{\mu}}_d^K(t-1) + 3\sigma\sqrt{\frac{\log t}{N_K(t-1)}}$$

which is equivalent to

$$3\sigma\sqrt{\frac{\log t}{N_i(t')}} - 3\sigma\sqrt{\frac{\log t}{N_K(t-1)}} \geq \hat{\hat{\mu}}_d^K(t-1) - \hat{\hat{\mu}}_d^i(t').$$

We have

$$\begin{aligned}
&\hat{\hat{\mu}}_d^K(t-1) - \hat{\hat{\mu}}_d^i(t') \\
&\geq \hat{\mu}_d^K(t-1) - \hat{\mu}_d^K(t') + 2\beta(N_K(t')) + \Delta_0 \\
&\geq -\beta(N_K(t-1)) - \beta(N_K(t')) + 2\beta(N_K(t')) + \Delta_0 \\
&\geq \Delta_0 > 0.
\end{aligned}$$

The first inequality is due to (9) and since Alice never attacks arm K we have $\hat{\hat{\mu}}^K(t-1) = \hat{\mu}^K(t-1)$. The second inequality holds by the definition of event E and the third inequality uses the fact that $\beta(n) = \sqrt{\frac{2\sigma^2}{n}\log\frac{\pi^2 K n^2}{3\delta}}$ is monotone decreasing in $n \geq 1$ if $K \geq \frac{3\delta e^2}{\pi^2}$ which can be shown by calculus.

Therefore, we have that

$$3\sigma\sqrt{\frac{\log t}{N_i(t')}} - 3\sigma\sqrt{\frac{\log t}{N_K(t-1)}} \geq \Delta_0 > 0, \tag{10}$$

and thus $N_K(t) = N_K(t-1) \geq N_i(t') + 1 = N_i(t)$.

Meanwhile, from (10) we get

$$3\sigma\sqrt{\frac{\log t}{N_i(t')}} > \Delta_0$$

and thus

$$N_i(t) = 1 + N_i(t') \leq 2 + \frac{9\sigma^2}{\Delta_0^2}\log t,$$

which completes the proof. □

Lemma 8.6. *Let us assume event E holds. Then the cumulative attack cost for any arm $i < K$ up to time step $t \geq 2K$, can be bounded as*

$$\sum_{\substack{a_s=i \\ s\leq t}} \alpha_s \leq \max_j N_j(t)(\Delta_j + \Delta_0 + 4\beta(N_j(t))).$$

Proof. Note that the right hand side is monotone increasing in t. It suffices to show that the result holds for t with $a_t = i$ essentially assuming that $N_K(t-1) = N_K(t)$.

By its definition, the cumulative cost satisfies

$$\begin{aligned}
\alpha_t &= \max\{\max_{j\in O_A^t, d} N_j(t)(\hat{z}_d^j - \bar{z}_d^j), 0\} \\
&= \max\{\max_{j\in O_A^t, d} N_j(t)\hat{\mu}_d^j(t) - \sum_{s=1}^{t-1}\alpha_s - N_j(t)\cdot(\hat{\mu}_d^K(t) - \\
&\quad 2\beta(N_K(t)) + \Delta_0), 0\} \\
&= \max\{\max_{j\in O_A^t, d} (N_j(t))(\hat{\mu}_d^j(t) - (\hat{\mu}_d^K(t) - \\
&\quad 2\beta(N_K(t)) - \Delta_0)) - \sum_{s=1}^{t-1}\alpha_s, 0\},
\end{aligned}$$

which implies by adding $\sum_{\substack{s=1 \\ a_s=i}}^{t-1}\alpha_s$ to both sides that

$$\sum_{\substack{s=1 \\ a_s=i}}^{t}\alpha_s \leq \max\{\max_{j\in O_A^t, d}\{N_j(t)(\hat{\mu}_d^j(t) - \hat{\mu}_d^K(t-1) + 2\beta(N_K(t-1)) + \Delta_0)\}, \sum_{\substack{s=1 \\ a_s=i}}^{t-1}\alpha_s\}. \tag{11}$$

Note that $\sum_{\substack{s=1 \\ a_s=i}}^{t-1}\alpha_s \leq \sum_{s=1}^{t-1}\alpha_s$. Since event E holds, we have

that

$$\begin{aligned}
&\sum_{\substack{s=1\\a_s=i}}^{t}\alpha_s\\
&\le \max\{\max_{j\in O_A^t,d} N_j(t)(\mu_d^j+\beta(N_j(t)))-\mu_d^K+\\
&\qquad \beta(N_K(t-1))+2\beta(N_K(t-1))+\Delta_0,\sum_{\substack{s=1\\a_s=i}}^{t-1}\alpha_s\}\\
&\le \max\{\max_{j\in O_A^t,d} N_j(t)(\mu_d^j-\mu_d^K+4\beta(N_j(t)))+\Delta_0,\sum_{\substack{s=1\\a_s=i}}^{t-1}\alpha_s\}\\
&= \max\{\max_{j\in O_A^t} N_j(t)(\Delta_j+\Delta_0+4\beta(N_j(t))),\sum_{\substack{s=1\\a_s=i}}^{t-1}\alpha_s\}\\
&\le \max_{j\in O_A^t} N_j(t)(\Delta_j+\Delta_0+4\beta(N_j(t)))
\end{aligned}$$

where the first inequality holds by the definition of event E, the second inequality holds as a result of Lemma 8.5 and the monotonicity of $\beta(\cdot)$ and the third inequality holds by the induction over $\sum_{s=1}^{t-1}\alpha_s$. The induction is with respect to all times $\bar{t}$ such that $a_{\bar{t}}=i$. The base case corresponds to $1\le\bar{t}\le 2K$ since we assume that initially every arm is pulled. If there exists $\bar{t}$, $K\le\bar{t}\le 2K$, such that $a_{\bar{t}}=i$, then this is the base case. Otherwise $1\le\bar{t}<K$ with $a_t=i$ and this is the base case. We also use the fact that $\sum_{s=1}^{2K}\alpha_s=0$.

□

Suppose event E holds. Lemma 8.5 proves the first half of the theorem.

For the second half, we observe that the total attack cost satisfies

$$\begin{aligned}
&\sum_{i<K}\sum_{\substack{a_s=i\\s\le T}}\alpha_s\\
&\le \sum_{i<K}\max_j N_j(T)(\Delta_j+\Delta_0+4\beta(N_j(T)))\\
&\le (K-1)\left(2+\frac{9\sigma^2}{\Delta_0^2}\log T\right)\cdot\\
&\qquad \max_j(\Delta_j+\Delta_0+4\beta(N_j(t)))\\
&\le (K-1)\left(2+\frac{9\sigma^2}{\Delta_0^2}\log T\right)\max_j(\Delta_j+\Delta_0+4\beta(2))\\
&= (K-1)\left(2+\frac{9\sigma^2}{\Delta_0^2}\log T\right)\max_j(\Delta_j+\Delta_0)+O^*(\log T)
\end{aligned}$$

where the first inequality holds by Lemma 8.6 and the second inequality again uses Lemma 8.5. The last inequality holds by the fact that $\beta(\cdot)$ is monotone decreasing.

According to Lemma 8.4, we have $P(E)\ge 1-D\delta$ and subsequently the results hold with probability at least $1-D\delta$.

□

8.2.7. Proof of Theorem 5.6

Proof of Theorem 5.6. We note that

$$\begin{aligned}
R_T &= \sum_{i=1}^{K} N_i(T)Dist(\mu^i,O)\\
&\ge N_K(T)\cdot Dist(\mu^K,O).
\end{aligned}$$

By construction of the instances aforementioned, for any arm $i\in O$, we have $\mu^K\prec\mu^i$ which leads to $Dist(\mu^K,O)>0$.

By Theorem 5.5 we derive $N_K(T)=T-\sum_{i<K}N_i(T)\ge T-(K-1)O^*(\log T)=O^*(T)$.

To conclude, the Pareto regret satisfies $R_T\ge N_K(t)\cdot Dist(\mu^K,O)=O^*(T)$ since $\mu^K\prec\mu^i$ implies $Dist(\mu^K,O)>0$. □

8.2.8. Proof of Theorem 5.7

Proof. The proof is two-fold. We first show that the two Pareto optimal fronts $\bar{O}'$ and O' can be quite close in a high probability sense and then show that the obtained reward vectors approach the mean vectors as T goes large by concentration inequalities.

We first note that the conditions in the theorem imply $\gamma\le\frac{1}{5}$ and $K=2$ is valid for Theorem 9. This implies that we can meet Assumption 4.

Let η be fixed with $0<\eta<1$. Since $\{r^{i,t}\}_{1\le t\le T}$ are i.i.d. sub-Gaussian distributed, the Chernoff-Hoeffding inequality implies

$$P(|\frac{1}{N_i(T)}\cdot\sum_{a_s=i}r^{i,s}-\mu^i|\not\preceq\gamma\cdot 1)\le 2D\exp(-\frac{\gamma^2}{2\sigma^2}). \tag{12}$$

In (12) we use the derivation from (8).

By choosing γ such that $2DK\exp(-\frac{\gamma^2}{2\sigma^2})\le\eta$ which is imposed in the theorem, we have

$$\begin{aligned}
&P(|\frac{1}{N_i(T)}\cdot\sum_{a_s=i}r^{i,s}-\mu^i|\preceq\gamma\cdot 1,\forall 1\le i\le K)\\
&\ge 1-\eta.
\end{aligned}\tag{13}$$

We denote the event as E_0 when using (13), i.e. $P(E_0)\ge 1-\eta$.

Likewise, note that

$$P(|\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}-\mu^i|\preceq\gamma\cdot 1,\forall 1\le i\le K)\ge 1-\eta \tag{14}$$

which can be shown by

$$\begin{aligned}
&P(|\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}-\mu^i|\preceq\gamma\cdot 1,\forall 1\le i\le K)\\
&=P(\forall i,\forall d:|\frac{1}{T}\cdot\sum_{t=1}^{T} r_d^{i,t}-\mu_d^i|<\gamma)\\
&=1-P(\exists i,\exists d:|\frac{1}{T}\cdot\sum_{t=1}^{T} r_d^{i,t}-\mu_d^i|>\gamma)\\
&\ge 1-\sum_{i=1}^{K}\sum_{d=1}^{D}P(|\frac{1}{T}\cdot\sum_{t=1}^{T} r_d^{i,t}-\mu_d^i|>\gamma)\\
&\ge 1-2DK\exp(-\frac{T\gamma^2}{2\sigma^2})\\
&\ge 1-\eta
\end{aligned}$$

where the first inequality is by the Bonferroni's inequality, the second inequality is by the Chernoff-Hoeffding inequality and the last one holds by the choice of γ. Note that if we use (14), we denote the event as E_1, i.e. $P(E_1)\ge 1-\eta$.

Meanwhile, by Assumption 4, $\mu^j-\mu^i\succeq 5\gamma\cdot 1$ holds for any arm i not in $\bar{O}_A$ and arm j in $\bar{O}_A$.

Note that O and $\bar{O}'$ only differs in whether $\frac{1}{N_i(T)}\sum_{a_t=i}\alpha_t$ is present. Since $K=2$, the two terms for attack cost are equivalent by noting that

$$\frac{1}{N_i(T)}\sum_{a_t=i}\alpha_t=\frac{1}{\sum_{j\neq K}N_j(T)}\cdot\sum_{a_t\neq K}\alpha_t. \tag{15}$$

For any arm $j\in\bar{O}'_A$ and arm i not in $\bar{O}'_A$, on event E_1 we derive

$$\begin{aligned}
&-\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}+\frac{1}{T}\cdot\sum_{t=1}^{T} r^{j,t}\\
&=-\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}+\mu^i-\mu^i+\\
&\quad\frac{1}{T}\cdot\sum_{t=1}^{T} r^{j,t}-\mu^j+\mu^j\\
&=-\mu^i+\mu^j-\\
&\quad(\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}-\mu^i)+(\frac{1}{T}\cdot\sum_{t=1}^{T} r^{j,t}-\mu^j)\\
&\succeq(5\gamma-2\gamma)\cdot 1=3\gamma\cdot 1>0.
\end{aligned} \tag{16}$$

where the last inequality is by the choice of i,j and (14).

This implies that $i\notin O'_A$ together with (15).

Similarly, for any arm $j\in O'_A$ and arm $i\notin O'_A$, on event E_1 we have

$$\begin{aligned}
&-\mu^j+\mu^i\\
&=-\frac{1}{T}\cdot\sum_{t=1}^{T} r^{j,t}+\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}-\\
&\quad(\frac{1}{T}\cdot\sum_{t=1}^{T} r^{i,t}-\mu^i)+(\frac{1}{T}\cdot\sum_{t=1}^{T} r^{j,t}-\mu^j)\\
&\preceq(0+\gamma+\gamma)\cdot 1=2\gamma\cdot 1
\end{aligned} \tag{17}$$

where the last inequality is by the choice of i,j and (14). Note that since $K=2$ we have $\sum_{t=1}^{T} r^{i,t}\preceq\sum_{t=1}^{T} r^{j,t}$.

If $i\in\bar{O}'_A$, then $j\notin\bar{O}'_A$ and $\mu^i>\mu^j+5\gamma\cdot 1$ since 1) there is an arm not in $\bar{O}'_A$ (by our construction $\bar{O}'_A$ contains at least one arm, arm K), and 2) Assumption 4 is valid. This contradicts (17) and thus implies that $i\notin\bar{O}'_A$.

Consequently, we established that on event E_1 we have $\bar{O}'_A=O'_A$.

Since $P(E_1)\ge 1-\eta$ we conclude

$$P(\bar{O}'_A=O'_A)\ge 1-\eta. \tag{18}$$

We further obtain by (14) and (15), on event E_1 for any i,

$$\begin{aligned}
&(\frac{1}{T}\sum_{t=1}^{T} r^{i,t}-\frac{1}{N_i(T)}\sum_{a_t=i}\alpha_t\cdot 1)-\gamma\\
&\quad\preceq\mu^i-\frac{1}{\sum_{j\neq K}N_j(T)}\cdot\sum_{a_t\neq K}\alpha_t\cdot 1\\
&\quad\preceq(\frac{1}{T}\sum_{t=1}^{T} r^{i,t}-\frac{1}{N_i(T)}\sum_{a_t=i}\alpha_t\cdot 1)+\gamma.
\end{aligned} \tag{19}$$

Therefore, by the definition of $\bar{O}'$ being defined as the Pareto optimal front of vectors $\mu^i-\frac{1}{\sum_{j\neq K}N_j(T)}\cdot\sum_{a_t\neq K}\alpha_t\cdot 1$ over all arm i and O' being over vectors $\frac{1}{T}\sum_{t=1}^{T} r^{i,t}-$

$\frac{1}{N_i(T)}\sum_{a_t=i}\alpha_t\cdot 1$, we have on event E_1 that

$$\begin{aligned}
&R'_T\\
&= T\cdot Dist(\frac{1}{T}\sum_{t=1}^{T} r^{a_t,t}\\
&\qquad - \frac{1}{\sum_{i\neq K}N_i(T)}\sum_{a_t\neq K}\alpha_t\cdot 1, O')\\
&\geq T\cdot Dist(\frac{1}{T}(N_K(T)\cdot\hat{\mu}^K+\sum_{i\neq K}N_i(T)\cdot\hat{\mu}^i)\\
&\qquad - \frac{1}{\sum_{i\neq K}N_i(T)}\sum_{a_t\neq K}\alpha_t\cdot 1, \bar{O}') - 2\gamma T. \quad (20)
\end{aligned}$$

The inequality follows from (18) and (19) and the following statement in Proposition 1.

Proposition 1 For any $\gamma_1>0, \gamma_2>0$ if ${O'}_A=\bar{O}'_A$ and $u^a\succeq v^a-\gamma_1\cdot 1$ for any $v^a\in\bar{O}', u^a\in O'$ where $a\in{O'}_A$, then for any e,$\bar{e}$ with $e\preceq\bar{e}+\gamma_2\cdot 1$ by Lemma 8.1 we have

$$\begin{aligned}
Dist(e,O') &= \min_d\max_{a\in{O'}_A}(u^a_d-e_d)\\
&= \min_d\max_{a\in\bar{O}'_A}(u^a_d-e_d)\\
&\geq \min_d\max_{a\in\bar{O}'_A}(v^a_d-e_d-\gamma_1)\\
&\geq \min_d\max_{a\in\bar{O}'_A}(v^a_d-\bar{e}_d)-\gamma_1-\gamma_2\\
&= Dist(\bar{e},\bar{O}')-\gamma_1-\gamma_2.
\end{aligned}$$

□

Note that $Dist(e,A)=Dist(e-z,\{a-z|a\in A\})$. On event $E_0\cap E_1$, we have that

$$\begin{aligned}
&R'_T\\
&\geq T\cdot Dist\left(\frac{1}{T}\cdot\left(N_K(T)\cdot\hat{\mu}^K+\sum_{i\neq K}N_i(T)\cdot\hat{\mu}^i\right),O\right)\\
&\qquad - 2\gamma T\\
&\geq T\cdot Dist\left(\frac{1}{T}\cdot\left(N_K(T)\cdot\mu^K+\sum_{i\neq K}N_i(T)\cdot\mu^i\right),O\right)\\
&\qquad - 3\gamma T\\
&\doteq T\mathcal{D}-3\gamma T.
\end{aligned}$$

where in the first inequality we use (15), the various definitions and the second inequality is a result of (13) by bounding μ_i with $\hat{\mu}+\gamma\cdot 1$ for any arm i.

By Theorem 5.5, we have that on event E

$$N_i(T)\leq O(logT) \quad (21)$$

and clearly $N_K(T)\leq T$.

Therefore, on events E, E_0, E_1 we have

$$\begin{aligned}
T\mathcal{D} &\geq T\cdot Dist(\frac{1}{T}\cdot(T\cdot\mu^K+\\
&\qquad (K-1)O(logT)\max_{i\neq K}\mu^i), O)\\
&\geq T\cdot Dist(\mu^K,O)-(K-1)O(logT) \quad (22)
\end{aligned}$$

where the last inequality uses the fact that $\max_i\mu^i\leq 1$.

By our construction, the distance from the reward vector of arm K to the Pareto optimal front O, $Dist(\mu^K,O)\geq 5\gamma$ since $K\notin O$. Therefore, on $E\cap E_0\cap E_1$ we have that

$$\begin{aligned}
R'_T &\geq T\mathcal{D}-3\gamma T\\
&\geq T\cdot 5\gamma-(K-1)O(logT)-3\gamma T\\
&= 2\gamma T-(K-1)\cdot O(logT).
\end{aligned}$$

Note that $P(E\cap E_0\cap E_1)\geq 1-P(E^c)-P((E_0\cap E_1)^c)=-D\delta+P(E_0\cap E_1)\geq -D\delta+1-P(E_0^c)-P(E_1^c)=1-D\delta-2\eta$. This completes the proof.

□

8.2.9. Proof of Theorem 5.8

Proof. We use the notation from the proof of Theorem 5.7. By the statement in (14), we obtain

$$P(|\frac{1}{T}\cdot\sum_{t=1}^{T}r^{i,t}-\mu^i|\preceq\gamma\cdot 1)\geq 1-\eta.$$

By Assumption 4, $\mu^j-\mu^i\succeq 5\gamma\cdot 1$ holds for any arm i not in $\bar{O}'_A$ and arm j in $\bar{O}'_A$.

For any arm $j\in\bar{O}'_A$ and arm i not in $\bar{O}'_A$, on event E_1 we have

$$-\frac{1}{T}\cdot\sum_{t=1}^{T}r^{i,t}+\frac{1}{T}\cdot\sum_{t=1}^{T}r^{j,t}\succ 0$$

by the result in (16).

Meanwhile, since $\bar{O}'$ and O' consider the same attack cost, this again implies that $i\notin{O'}_A$.

Similarly, for any arm $j\in{O'}_A$ and arm $i\notin{O'}_A$, on event E_1 we have arm $i\notin\bar{O}'_A$.

Consequently on event E_1 we get

$${O'}_A=\bar{O}'_A$$

and

$$\begin{aligned}&(\frac{1}{T}\sum_{t=1}^{T} r^{i,t} - \frac{1}{T}\sum_{t=1}^{T}\bar{\alpha}_t^i \cdot 1) - \gamma \\ &\preceq \mu^i - \frac{1}{T}\sum_{t=1}^{T}\bar{\alpha}_t^i \cdot 1 \\ &\preceq (\frac{1}{T}\sum_{t=1}^{T} r^{i,t} - \frac{1}{T}\sum_{t=1}^{T}\bar{\alpha}_t^i \cdot 1) + \gamma.\end{aligned}$$

Therefore, by the definition of $\bar{O}'$ being defined as the Pareto front of vectors $\mu^i - \frac{1}{T}\sum_{t=1}^{T}\bar{\alpha}_t^i \cdot 1$, whereas O' is over $\frac{1}{T}(\sum_{t=1}^{T} r^{i,t} - \sum_{t=1}^{T}\bar{\alpha}_t^i \cdot 1)$, we have on event $E_0 \cap E_1$ that

$$R'_T \geq T \cdot Dist(\frac{1}{T} \cdot \left(N_K(T)\cdot\mu^K + \sum_{i\neq K} N_i(T)\cdot\mu^i\right) - \frac{1}{T}\sum_{t=1}^{T}\bar{\alpha}_t^{a_t}\cdot 1, \bar{O}') - 2\gamma T. \quad (23)$$

This derivation follows (20), (13) and relies on Proposition 1.

Consider the counterfactual attack cost $\bar{\alpha}_t^i$ on arm i in the definition of $\bar{O}'$. Formally, according to the definition of $\bar{\alpha}_t^i$, it reads explicitly as

$$\bar{\alpha}_t^i = \max\{\max_d N_i(t)(\hat{z}_d^i - \bar{z}_d^i), 0\}.$$

We observe that

$$\begin{aligned}\alpha_t &= \max\{\max_{j\in O_A^t, d} N_j(t)(\hat{z}_d^j - \bar{z}_d^j), 0\} \\ &\geq \bar{\alpha}_t^{a_t}\end{aligned}$$

since the chosen arm by Bob $a_t \in O_A^t$.

Therefore, by the results in Lemma 8.6 and Lemma 8.5 on event E, we have

$$\begin{aligned}&\sum_{t=1}^{T}\bar{\alpha}_t^{a_t} \leq \sum_{t=1}^{T}\alpha_t \\ &\leq (K-1)\max_{j<K}\sum_{\substack{a_s=j \\ s\leq T}}\alpha_s \\ &\leq (K-1)\max_j N_j(T)(\Delta_j + \Delta_0 + 4\beta(N_j(T))) \\ &\leq (K-1)\left(2 + \frac{9\sigma^2}{\Delta_0^2}\log T\right)\max_j(\Delta_j + \Delta_0 + 4\beta(2)) \\ &= O(\log T).\end{aligned}$$

This leads to $\frac{1}{T}\cdot\sum_{t=1}^{T}\bar{\alpha}_t^{a_t} \to 0$ as $T \to \infty$.

Meanwhile for any $i \neq K$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\bar{\alpha}_t^i \\ &= \sum_{t=1}^{T}\max\{\max_d N_i(t)(\hat{z}_d^i - \bar{z}_d^i), 0\} \\ &\leq \sum_{t=1}^{T}\max\{\max_{j\in O_A^t, d} N_i(t)(\hat{z}_d^j - \bar{z}_d^j), 0\} \\ &= \sum_{t=1}^{T}\alpha_t \leq O(\log T).\end{aligned}$$

As a result, we further obtain $\frac{1}{T}\cdot\sum_{t=1}^{T}\bar{\alpha}_t^i \to 0$ as $T \to \infty$.

Suppose for $T \geq T(\gamma)$ we have $\frac{1}{T}\cdot\sum_{t=1}^{T}\alpha_t^i \leq \gamma$ and $\frac{1}{T}\cdot\sum_{t=1}^{T}\alpha_t^{a_t} \leq \gamma$.

From (23) we further obtain by using Proposition 1

$$\begin{aligned}&R'_T \\ &\geq T \cdot Dist\left(\frac{1}{T}\cdot\left(N_K(T)\cdot\mu^K + \sum_{i\neq K}N_i(T)\cdot\mu^i\right), O\right) \\ &\quad - 2\zeta\cdot T - 2\gamma T \\ &\doteq T\mathcal{D} - 2\gamma T - 2\gamma T.\end{aligned}$$

By the result in (22), we further have that on $E \cap E_0 \cap E_1$

$$T\mathcal{D} \geq T \cdot Dist(\mu^K, O) - (K-1)\cdot O(logT)$$

Since by assumption we have $Dist(\mu^K, O) \geq 5\gamma$, we conclude that for $T \geq T(\gamma)$ with probability at least $1-2\eta-D\delta$, R'_T is of order T by noting that

$$\begin{aligned}&R'_T \geq T\mathcal{D} - 2\gamma T - 2\gamma T \\ &\geq T\cdot 5\gamma - (K-1)\cdot O(logT) - 2\gamma T - 2\gamma T \\ &= \gamma T - (K-1)\cdot O(logT).\end{aligned}$$

□