Problem-dependent Regret for Lexicographic Multi-Armed Bandits with Adversarial Corruptions

Bo Xue[1,2] , **Xi Lin**[1,2] , **Yuanyu Wan**[3] and **Qingfu Zhang**[1,2,*]
[1]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[2]The City University of Hong Kong Shenzhen Research Institute, Shenzhen, China
[3]School of Software Technology, Zhejiang University, Ningbo, China
{boxue4-c, xi.lin}@my.cityu.edu.hk, wanyy@zju.edu.cn, qingfu.zhang@cityu.edu.hk

Abstract

This paper studies lexicographic multi-armed bandits (MAB), where after selecting an arm, the agent observes a reward vector including multiple objectives, each with a different level of importance. Although previous literature has proposed the algorithm for lexicographic MAB, their algorithm suffers from several limitations: (1) it exhibits poor adversarial robustness due to its reliance on stochastic rewards, (2) its regret bound is suboptimal compared to single-objective counterparts, and (3) the regret bound does not adapt to specific problem instances. To address these limitations, we study lexicographic MAB with adversarial corruptions, where an adversary might corrupt the stochastic rewards with a corruption budget of C. First, when the value of C is known, we propose an algorithm achieving a problem-dependent regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\log T}{\Delta^i(a)}+C\right)\right)$ for the i-th objective ($i\in[M]$), where $\Delta^i(a)$ is the reward gap for arm a on the i-th objective, and M is the number of objectives. In the purely stochastic setting ($C=0$), this regret bound approaches optimality. Second, we introduce another algorithm that does not require value of C but incurs a less favorable regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\gamma_T}{\Delta^i(a)}+\gamma_T\right)\right)$ for the i-th objective, where $\gamma_T=O((\log T)^2+KC(\log T)^2)$. Finally, we conduct experiments on both synthetic and real-world datasets to verify the effectiveness of our algorithms.

1 Introduction

Multi-armed bandits (MAB) has emerged as a prominent framework in the field of sequential decision-making [Robbins, 1952], where at each round, an agent first chooses one of K arms and then receives a reward related to the chosen arm. The goal of the agent is to maximize the cumulative rewards over T rounds. MAB has found applications in various domains, including online advertising [Schwartz *et al.*, 2017], clinical trials [Villar *et al.*, 2015], and resource allocation [Khansa *et al.*, 2021]. Despite its power, many real-world applications encounter the trade-off between different objectives. For example, supply chain management aims to maximize revenue while minimizing costs [Trisna *et al.*, 2016], and recommender systems strive to maximize the user engagement while ensuring the fairness [Wang *et al.*, 2023]. These applications have motivated the development of multi-objective bandits, where the observed reward is a vector containing multiple objectives [Hüyük and Tekin, 2021; Groetzner and Werner, 2022; Xu and Klabjan, 2023; Cai *et al.*, 2023; Cheng *et al.*, 2024; Xue *et al.*, 2025].

Most research on multi-objective bandits employs Pareto regret to evaluate algorithms, which assumes all objectives carry equal importance [Turgay *et al.*, 2018; Lu *et al.*, 2019; Cai *et al.*, 2023]. However, some real-world applications require different levels of importance among objectives. For instance, in radiation treatment for cancer patients, target coverage takes precedence over proximity to organs at risk [Jee *et al.*, 2007]. Similarly, water resource planning prioritizes objectives such as flood protection, irrigation supply shortages, and electricity generation [Weber *et al.*, 2002]. Moreover, Theorem 4.1 in Xu and Klabjan [2023] states that Pareto regret is lower than the individual regret of any objective. Therefore, optimizing any single objective among the multiple objectives achieves the optimal Pareto regret bound, while the other objectives still suffer linear regret bounds of $O(T)$.

To address this issue, lexicographic order is proposed [Ehrgott, 2005], which distinguishes objectives through their importance. Precisely, in a lexicographic bandit problem with M objectives, the i-th objective is more important than the j-th objective if $i<j$ and $i,j\in[M]$[1]. As the most related work, Hüyük and Tekin [2021] investigated multi-objective multi-armed bandits (MOMAB) under lexicographic order and defined a metric called priority-based regret. Precisely, let $\mathbb{I}(\cdot)$ be the indicator function, and $\mu^i(a)\in\mathbb{R}$ be the expected reward of arm $a\in[K]$ on the objective $i\in[M]$. The priority-based regret for the objective $i\in[M]$ is

$$\widehat{R}^i(T)=\sum_{t=1}^{T}\Delta^i(a_t)\cdot\mathbb{I}(\mu^j(a_t)=\mu^j(a_*),\forall j\in[i-1]),\quad (1)$$

where $\Delta^i(a)=\mu^i(a_*)-\mu^i(a)$ is the expected reward gap, a_t

[*] Qingfu Zhang is the corresponding author.

[1] For any positive integer N, $[N]$ denotes the set $\{1,2,\ldots,N\}$.

is the arm chosen at the t-th round, and a_* is the lexicographic optimal arm (which will be defined in Section 3). Based on this metric, Hüyük and Tekin [2021] developed an algorithm whose regret bound is $\widetilde{O}((KT)^{2/3})$.

Although this work [Hüyük and Tekin, 2021] establishes a foundational framework for addressing the lexicographic bandit problem, it has four limitations: (1) It assumes that rewards are stochastic, which may be violated in some practical scenarios, such as click fraud in pay-per-click online advertising [Wilbur and Zhu, 2009] and malicious reviews in recommendation systems [Lykouris *et al.*, 2018]. (2) The regret bound $\widetilde{O}((KT)^{2/3})$ is suboptimal, as the problem-dependent lower bound and minimax lower bound for single-objective MAB are $\Omega\left(\sum_{\Delta(a)>0} \frac{\log T}{\Delta(a)}\right)$ and $\Omega(\sqrt{KT})$, respectively [Bubeck and Cesa-Bianchi, 2012]. (3) Its regret bound cannot adapt to specific problem instances. (4) The priority-based regret is inaccurate in certain cases. For example, if $\mu^1(a_t) < \mu^1(a_*)$, then $\mathbb{I}(\mu^j(a_t) = \mu^j(a_*), \forall j \in [i-1]) = 0$ for $i \geq 2$. Therefore, the instantaneous gap $\Delta^i(a_t)$ for $i \geq 2$ is ignored because $\Delta^i(a_t) \cdot \mathbb{I}(\mu^j(a_t) = \mu^j(a_*), \forall j \in [i-1]) = 0$.

To address these limitations, we study the lexicographic MOMAB with adversarial corruptions, whose most rewards are stochastic, but a small fraction can be contaminated by an adversary [Lykouris *et al.*, 2018]. With a budget of corruptions C, our main contributions are summarized as follows:

- We adopt a more accurate metric to evaluate lexicographic MAB algorithms, which is a natural extension of the regret in single-objective bandits [Auer *et al.*, 2002], i.e.,
$$R^i(T) = \sum_{t=1}^{T} \Delta^i(a_t), i \in [M]. \tag{2}$$
- If C is known, we propose an algorithm that enjoys a regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\log T}{\Delta^i(a)} + C\right)\right)$ for the i-th objective, $i \in [M]$. In the stochastic setting ($C = 0$), this regret bound becomes $O\left(\sum_{\Delta^i(a)>0} \frac{\log T}{\Delta^i(a)}\right)$, which matches the lower regret bound of MAB [Bubeck and Cesa-Bianchi, 2012] and improves upon the existing regret bound $\widetilde{O}((KT)^{2/3})$ [Hüyük and Tekin, 2021].
- If C is unknown, we employ the multi-instance technique [Lykouris *et al.*, 2018] to design a new algorithm that has a regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\gamma_T}{\Delta^i(a)} + \gamma_T\right)\right)$ for the i-th objective, where $\gamma_T = O((\log T)^2 + KC(\log T)^2)$. When $C = 0$, this regret bound becomes $O\left(\sum_{\Delta^i(a)>0}\left(\frac{(\log T)^2}{\Delta^i(a)} + (\log T)^2\right)\right)$, which also improves the existing lexicographic bandit algorithm [Hüyük and Tekin, 2021].
- We conduct two sets of experiments to validate our theoretical findings. The first set consists of synthetic experiments, while the second utilizes a real-world dataset related to COVID-19 vaccines. Results from both sets of experiments demonstrate that our algorithms can effectively optimize multiple objectives simultaneously.

2 Related Work

In this section, we review the research on corruption-tolerant bandits and multi-objective bandits.

2.1 Corruption-tolerant Bandits

Lykouris *et al.* [2018] first introduced this new bandit model called MAB with adversarial corruptions, and proposed two basic ideas for this model. If the budget C is known, Lykouris *et al.* [2018] utilized enlarged confidence intervals to achieve a regret bound of $O\left(\sum_{\Delta(a)>0} \frac{\log T + C}{\Delta(a)}\right)$.[2] If C is unknown, Lykouris *et al.* [2018] proposed the multi-instance technique which dynamically adapts to the corruptions and achieved a regret bound of $O\left(\sum_{\Delta(a)>0} \frac{KC(\log T)^2 + (\log T)^2}{\Delta(a)}\right)$. Later, Gupta *et al.* [2019] further improved this regret bound to $O\left(KC + \sum_{\Delta(a)>0} \frac{\log(T)\log(\log T)}{\Delta(a)}\right)$ by randomly selecting arms from a specially designed distribution. Gupta *et al.* [2019] also established a lower bound that exhibits a linear relationship with the budget C. Subsequent research has advanced the corruption-tolerant MAB into various directions, including stochastic linear bandits [He *et al.*, 2022; Ding *et al.*, 2022], Lipschitz bandits [Kang *et al.*, 2023], graph bandits [Lu *et al.*, 2021], combinatorial bandits [Balasubramanian *et al.*, 2024] and Gaussian bandits [Bogunovic *et al.*, 2020]. Beyond the budget-bounded setting, several papers have focused on a similar scenario in which an adversary attacks with a certain probability at each round, and these attack values can be unbounded [Altschuler *et al.*, 2019; Guan *et al.*, 2020; Mukherjee *et al.*, 2021].

2.2 Multi-objective Bandits

Drugan and Nowe [2013] initially formalized the MOMAB problem and proposed two algorithms that exhibit regret bounds of $O\left(\sum_{\Delta(a)>0}\left(\frac{\log T}{\Delta(a)} + \Delta(a)\right)\right)$ for scalarized regret and Pareto regret, respectively. Turgay *et al.* [2018] devoted into the multi-objective contextual bandits and proposed a zooming-based algorithm that achieves a Pareto regret bound of $\widetilde{O}(T^{(d_p+1)/(d_p+2)})$, where d_p is Pareto zooming dimension. Lu *et al.* [2019] explored the multi-objective generalized linear bandits and provided a regret bound of $\widetilde{O}(\sqrt{T})$. Chowdhury and Gopalan [2021] studied multi-objective bandit learning from the perspective of nonparametric Bayesian optimization. Another line of research is the Pareto set identification, whose goal is to analyze the cost of identifying all Pareto optimal arms [Auer *et al.*, 2016; Ararat and Tekin, 2023; Kone *et al.*, 2023]. Tekin and Turgay [2018] initially examined lexicographic contextual bandits, establishing a regret bound of $\widetilde{O}(T^{(d_c+2)/(d_c+3)})$, where d_c is the dimension of context information. However, their study was limited to scenarios with two objectives. Xue *et al.* [2024] later extended the number of objectives beyond two in the Lipschitz bandit setting, deriving a regret bound of $\widetilde{O}(T^{(d_z^i+1)/(d_z^i+2)})$ for the i-th objective, where d_z^i is the zooming dimension and $i \in [M]$. The most related work is

[2] For the single-objective setting, $\Delta^1(a)$ is simplified as $\Delta(a)$.

by Hüyük and Tekin [2021], which studied the lexicographic MOMAB and achieved a regret bound of $\widetilde{O}((KT)^{2/3})$.

Existing algorithms for lexicographic bandits can only handle stochastic rewards, which lacks robustness against adversarial attacks. Meanwhile, their regret bounds fail to adapt to specific problem instances. Therefore, we propose new algorithms for lexicographic MOMAB with adversarial corruptions and establish problem-dependent regret bounds.

3 Preliminaries

This paper studies lexicographic MOMAB with adversarial corruptions. Let M be the number of objectives and K be the number of arms. At each round $t = 1, 2, \ldots, T$, the formal protocol between the agent and adversary is as follows:

1. The environment assigns each arm $a \in [K]$ a stochastic reward $r_t(a) = [r_t^1(a), \ldots, r_t^M(a)] \in \mathbb{R}^M$, whose expectation is $\mu(a) = [\mu^1(a), \ldots, \mu^M(a)]$. For any objective $i \in [M]$ and any arm $a \in [K]$, $\mu^i(a)$ is bounded in $[0, 1]$ and $r_t^i(a)$ satisfies the 1-sub-Gaussian property, i.e.,
$$\mathrm{E}\left[e^{\alpha\left(r_t^i(a)-\mu^i(a)\right)}\right] \leq e^{\alpha^2/2}, \forall \alpha \in \mathbb{R}. \tag{3}$$
2. The adversary observes the stochastic reward of any arm $a \in [K]$ and corrupts it to $\tilde{r}_t(a) = [\tilde{r}_t^1(a), \ldots, \tilde{r}_t^M(a)]$, where the corruptions are bounded, such that
$$|\tilde{r}_t^i(a) - r_t^i(a)| \leq 1, \forall i \in [M], a \in [K]. \tag{4}$$
3. The agent chooses an arm a_t and then observes the corrupted reward vector $\tilde{r}_t(a_t)$.

Following the previous studies on the corruption-tolerant bandits [Lykouris *et al.*, 2018; Gupta *et al.*, 2019], we assume the total corruptions of each objective are bounded by C, i.e.,

$$\sum_{t=1}^{T} \max_{a \in [K]} |\tilde{r}_t^i(a) - r_t^i(a)| \leq C, \forall i \in [M]. \tag{5}$$

Next, we introduce the lexicographic order so as to compare different arms by their expected rewards.

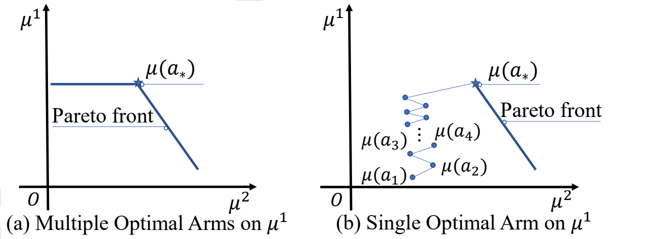
Definition 1 (**Lexicographic Order**). *Let $u, v \in \mathbb{R}^M$ be two vectors. u is said to lexicographically dominate v if and only if there exists some $i^* \in [M]$, such that $u^i = v^i$ for any $i \in [i^* - 1]$ and $u^{i^*} > v^{i^*}$.*

Lexicographic order is a total order allowing the comparison of any two vectors, thereby deciding the lexicographic optimal arm.

Definition 2 (**Lexicographic Optimal Arm**). *An arm a_* is lexicographic optimal if and only if its expected reward is not lexicographically dominated by that of any other arms in $[K]$.*

Finally, we introduce a concept termed local trade-off, which is similar to the established notion called global trade-off [Miettinen, 1999, Definition 2.8.5].

Definition 3 (**Local Trade-off**). *A positive real number λ_0 is the local trade-off parameter of a MOMAB problem if and only if it is the smallest $\lambda > 0$ that satisfies the condition: for any $i \geq 2$ and $a \in [K]$,*
$$\mu^i(a) - \mu^i(a_*) \leq \lambda \cdot \max_{j \in [i-1]} \{\mu^j(a_*) - \mu^j(a)\}. \tag{6}$$



(a) Multiple Optimal Arms on μ^1 (b) Single Optimal Arm on μ^1

Figure 1: **(a)** There exist multiple optimal arms for the first objective, thus a lexicographic bandit algorithm is necessary to identify the lexicographic optimal arm a_*. **(b)** There exists only one optimal arm for the first objective, thus running a single-objective bandit algorithm on the first objective is sufficient to determine the arm a_*. However, focusing on the first objective only cannot optimize the second objective.

The trade-off parameter λ_0 is a ratio indicating when transitioning from the lexicographic optimal arm a_* to other arms, how much the value of the i-th objective will increase per unit decrease in the preceding $i - 1$ objectives. Due to the inherent conflicts among different objectives, employing the information about trade-offs is common in multi-objective optimization [Kaliszewski, 2000; Keeney, 2002; Nowak and Trzaskalik, 2022].

4 Effective Scenarios of Lexicographic Bandit Algorithms

Before introducing our algorithms, we present two typical lexicographic bandit problems in Figure 1(a) and Figure 1(b), highlighting scenarios where our algorithms are effective.

In Figure 1(a), multiple arms attain the maximum expected reward on the first objective. Therefore, executing a single-objective bandit algorithm based solely on the rewards of the first objective cannot identify the lexicographic optimal arm a_*, and taking a lexicographic bandit algorithm is necessary.

In contrast, Figure 1(b) depicts a scenario where only one arm a_* attains the maximum expected reward on the first objective. Here, identifying the optimal arm for the first objective is equivalent to identifying the lexicographic optimal arm a_*. This raises a natural question:

Can a single-objective bandit algorithm replace lexicographic bandit algorithms in the scenario of Figure 1(b)?

The answer is **no**. As depicted in Figure 1(b), applying a single-objective bandit algorithm, such as UCB [Auer, 2002], to the first objective yields a sequence of chosen arms $\{a_1, a_2, \ldots, a_t, \ldots\}$. By the theoretical guarantee of UCB, the sequence of expected rewards for the first objective, $\{\mu^1(a_1), \mu^1(a_2), \ldots, \mu^1(a_t), \ldots\}$, converges to the optimal expected reward $\mu^1(a_*)$, ensuring a sublinear regret bound for the first objective, as shown in Figure 1(b). However, the sequence of expected rewards for the second objective, $\{\mu^2(a_1), \mu^2(a_2), \ldots, \mu^2(a_t), \ldots\}$, may deviate significantly from the optimal value $\mu^2(a_*)$, potentially resulting in linear regret for the second objective. Therefore, single-objective algorithms cannot optimize multiple objectives simultaneously in the scenario like Figure 1(b), thus cannot replace the lexicographic bandit algorithms.

Algorithm 1 Corruption-tolerant Multi-objective Bandits with Known Budget (CMOB-KB)

```
Input: $\delta \in (0,1)$, $\lambda \geq \lambda_0$
Initialize empirical mean rewards $\hat{\mu}^i(a) = 0$, counters $n(a) = 0$, and confidence terms $w(a) = +\infty$ for $i \in [M]$ and $a \in [K]$
Initialize exploration index $s = 1$ and arm set $\mathcal{A}_s = [K]$
for $t = 1, 2, \ldots$ do
    Invoke Algorithm 2 to select an arm, update the exploration index and candidate arm set: $a_t, s, \mathcal{A}_s = \text{AELO}\left(s, \mathcal{A}_s, \{\hat{\mu}^i(a), w(a)\}_{a \in \mathcal{A}_s}^{i \in [M]}\right)$
    Play the arm $a_t$ and observe the reward vector $\tilde{r}_t(a_t)$
    Update the empirical rewards and counter by Eq. (8)
    Update the confidence term by Eq. (9)
end for
```

Algorithm 2 AELO

```
Input: $s, \mathcal{A}_s, \{\hat{\mu}^i(a), w(a)\}_{a \in \mathcal{A}_s}^{i \in [M]}$
repeat
    if $w(a) > 2^{-s}$ for some $a \in \mathcal{A}_s$ then
        Choose the arm $a_t = \text{argmin}_{a \in \mathcal{A}_s} n(a)$
    else
        Initialize the arm set $\mathcal{A}_s^0 = \mathcal{A}_s$
        for $i = 1, 2, \ldots, M$ do
            $\hat{a}_t^i = \text{argmax}_{a \in \mathcal{A}_s^{i-1}} \hat{\mu}^i(a) + w(a)$
            $\mathcal{A}_s^i = \{a \in \mathcal{A}_s^{i-1} | \hat{\mu}^i(a) + w(a) \geq \hat{\mu}^i(\hat{a}_t^i) + w(\hat{a}_t^i) - (2 + 4\lambda + \ldots + 4\lambda^{i-1}) \cdot 2^{-s}\}$
        end for
        Update $\mathcal{A}_{s+1} = \mathcal{A}_s^M$ and $s = s + 1$
    end if
until an arm $a_t$ is chosen
Return $a_t$, $s$ and $\mathcal{A}_s$
```

5 Algorithms

In this section, we first propose a budget-dependent algorithm, and then present a budget-free algorithm to remove the requirement on the budget value C.

5.1 Budget-dependent Method: CMOB-KB

As a warm-up, we introduce a straightforward algorithm that requires the value of budget C, called Corruption-tolerant Multi-objective Bandits with Known Budget (CMOB-KB).

At the start, CMOB-KB initializes the empirical mean rewards $\hat{\mu}^i(a)$ and the counter $n(a)$ for any objective $i \in [M]$ and any arm $a \in [K]$ as zero, where the counter is used to track the number of times each arm is played. The confidence term $w(a)$ for any arm $a \in [K]$ is initialized as infinity. In addition, CMOB-KB initializes an exploration index $s = 1$ and a candidate arm set $\mathcal{A}_s = [K]$. With these initializations, CMOB-KB is ready to start decision-making. To balance the trade-off across different objectives, we design a novel decision-making method called Arm Elimination under Lexicographic Ordering (AELO), as outlined in Algorithm 2.

AELO is a repeated-until loop which iteratively eliminates arms in $\mathcal{A}_s$ until an arm is chosen. **If** there is an arm $a \in \mathcal{A}_s$ whose confidence term $w(a)$ is greater than 2^{-s}, AELO chooses the arm with the fewest number of plays, i.e., $a_t = \text{argmin}_{a \in \mathcal{A}_s} n(a)$, and terminates the loop. **If** no arm satisfies this condition, AELO sequentially eliminates arms from the first objective to the last objective so as to balance the exploration and exploitation across different objectives.

Let $\mathcal{A}_s^0 = \mathcal{A}_s$ be the initialized candidate arms. For the objective $i \in [M]$, AELO first selects the arm $\hat{a}_t^i$ from $\mathcal{A}_s^{i-1}$ that maximizes the upper confidence bound of this objective, i.e., $\hat{a}_t^i = \text{argmax}_{a \in \mathcal{A}_s^{i-1}} \hat{\mu}^i(a) + w(a)$. Then, AELO keeps the arms whose upper confidence bound is greater than or equal to the upper confidence bound of $\hat{a}_t^i$ minus a term depending on the objective order and the exploration index, i.e.,

$$\begin{aligned} \mathcal{A}_s^i = \{&a \in \mathcal{A}_s^{i-1} | \hat{\mu}^i(a) + w(a) \geq \hat{\mu}^i(\hat{a}_t^i) + w(\hat{a}_t^i) \\ &- (2 + 4\lambda + \ldots + 4\lambda^{i-1}) \cdot 2^{-s}\}. \end{aligned} \tag{7}$$

This step eliminates arms that are less promising. After eliminating arms on the last objective M, AELO sets $\mathcal{A}_{s+1}$ as $\mathcal{A}_s^M$ and s as $s + 1$. AELO repeats the above steps until an arm is chosen. Finally, AELO returns the chosen arm a_t, the updated index s and arm set $\mathcal{A}_s$.

Once CMOB-KB obtains the arm returned by AELO, it plays a_t and receives the reward $\tilde{r}_t(a_t)$. Then, CMOB-KB updates the empirical mean rewards of all objectives and the counter $n(a_t)$, i.e.,

$$\hat{\mu}^i(a_t) = \frac{n(a_t) \cdot \hat{\mu}^i(a_t) + \tilde{r}_t^i(a_t)}{n(a_t) + 1}, n(a_t) = n(a_t) + 1. \tag{8}$$

Finally, CMOB-KB computes $w(a_t)$ based on the updated $n(a_t)$, i.e.,

$$w(a_t) = \sqrt{\frac{\alpha(a_t)}{n(a_t)}} + \frac{C}{n(a_t)} \tag{9}$$

where $\alpha(a_t) = 4\log(4MKn(a_t)/\delta)$. After all these updates, CMOB-KB has finished the current trial and is prepared to make the next decision.

We provide the following theorem for CMOB-KB.

Theorem 1. *Suppose* (3), (4) *and* (5) *hold. If CMOB-KB is run with $\lambda \geq \lambda_0$, then with probability at least $1 - \delta$, for any objective $i \in [M]$, its regret satisfies*

$$R^i(T) \leq \sum_{\Delta^i(a) > 0} \frac{128\alpha_T(\Lambda^i(\lambda))^2}{\Delta^i(a)} + 32\Lambda^i(\lambda)C$$

where $\alpha_T = 4\ln(4MKT/\delta)$ and $\Lambda^i(\lambda) = 1 + \lambda + \ldots + \lambda^{i-1}$.

Remark 1. Theorem 1 states that CMOB-KB achieves a regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\alpha_T(\Lambda^i(\lambda))^2}{\Delta^i(a)} + \Lambda^i(\lambda)C\right)\right)$. Notably, $\Lambda^1(\lambda) = 1$, thus the regret bound of the first objective is $O\left(\sum_{\Delta^1(a)>0}\left(\frac{\alpha_T}{\Delta^1(a)} + C\right)\right)$, which aligns with the regret bound of the single-objective corruption-tolerant bandit algorithm [Lykouris *et al.*, 2018]. Therefore, CMOB-KB does not degrade the performance of the most important objective ($i = 1$) when optimizing multiple objectives simultaneously. Although subsequent objectives may have higher regrets, it is acceptable for lower-priority objectives to experience greater

Algorithm 3 Corruption-tolerant multi-objective Bandits with Unknown Budget (CMOB-UB)

```
Input: $\delta \in (0,1), \lambda \geq \lambda_0, T \in \mathbb{N}_+$
Compute instance number $L = \lceil \log_2 T \rceil$
Initialize empirical mean rewards $\hat{\mu}^i_\ell(a) = 0$, counters $n_\ell(a) = 0$, and confidence terms $w_\ell(a) = +\infty$ for $i \in [M], a \in [K]$ and $\ell \in [L]$
Initialize exploration index $s_\ell = 1$ and candidate arm set $\mathcal{A}_\ell = [K]$ for $\ell \in [L]$
for $t = 1, 2, \ldots, T$ do
    Sample instance $\ell_t \in [L]$ with probability $2^{-\ell_t}$, with the remaining probability sampling $\ell_t = 1$
    if $\mathcal{A}_{\ell_t} \neq \emptyset$ then
        Invoke Algorithm 2 to select an arm, update exploration index and candidate arm set: $a_t, s_{\ell_t}, \mathcal{A}_{s_{\ell_t}} = \text{AELO}\left(s_{\ell_t}, \mathcal{A}_{s_{\ell_t}}, \{\hat{\mu}^i_{\ell_t}(a), w_{\ell_t}(a)\}^{i\in[M]}_{a\in\mathcal{A}_{s_{\ell_t}}}\right)$
        Update the arm set for previous instances: for $\ell < \ell_t$, $\mathcal{A}_{s_\ell} = \mathcal{A}_{s_\ell} \cap \mathcal{A}_{s_{\ell_t}}$
        Play the arm $a_t$ and observe the rewards $\tilde{r}_t(a_t)$
        Update empirical rewards and counter by Eq. (10)
        Update the confidence term by Eq. (11)
    else
        Find the minimum instance $\ell'_t \geq \ell_t$ such that $\mathcal{A}_{\ell'_t} \neq \emptyset$ and randomly play an arm from $\mathcal{A}_{\ell'_t}$
    end if
end for
```

regrets. This effect is typically modest given that most multi-objective problems involve two or three objectives [Deb and Jain, 2014; Li *et al.*, 2015].

Remark 2. In the setting of stochastic rewards ($C = 0$), the above regret bound reduces to $O\left(\sum_{\Delta^i(a)>0} \frac{\alpha_T(\Lambda^i(\lambda))^2}{\Delta^i(a)}\right)$, which matches the problem-dependent lower bound of single-objective bandits in terms of $\Delta^i(a)$ [Bubeck and Cesa-Bianchi, 2012]. This is a considerable improvement since the existing regret bound of lexicographic MOMAB work is $\widetilde{O}((KT)^{2/3})$ [Hüyük and Tekin, 2021], which deviates from the minimax regret bound $\widetilde{O}(\sqrt{KT})$ of the single-objective bandit algorithm [Bubeck and Cesa-Bianchi, 2012]. Furthermore, we adopt the general regret (2) as the metric, which is more accurate than the priority-based regret (1) used in [Hüyük and Tekin, 2021], as discussed in Section 1.

Remark 3. Regarding the budget value C, Gupta *et al.* [2019] established a budget-dependent lower bound $\Omega(C)$, which indicates that our regret bound is optimal concerning the budget C. Meanwhile, CMOB-KB is an anytime algorithm which does not take T as input,distinguishing it from existing methods [Lykouris *et al.*, 2018; Gupta *et al.*, 2019].

5.2 Budget-free Method: CMOB-UB

To remove the dependence on C, a commonly used approach is the multi-instance technique [Lykouris *et al.*, 2018; Kang *et al.*, 2023], which involves constructing multiple corruption-tolerant instances like CMOB-KB first, and then randomly picking an instance to run in each round. Since each instance has a different tolerance level against corruption and the more resilient one is less likely to be chosen, this approach enables automatic adaptation to the unknown value C. Leveraging this technique, we design a budget-free algorithm called Corruption-tolerant Multi-objective Bandits with Unknown Budget (CMOB-UB).

At the start, CMOB-UB calculates $L = \lceil \log_2 T \rceil$, which is the number of instances. Afterward, CMOB-UB initializes the empirical mean rewards $\hat{\mu}^i_\ell(a)$, the counters $n_\ell(a)$, and the confidence terms $w_\ell(a)$ for each objective $i \in [M]$, arm $a \in [K]$, and instance $\ell \in [L]$. Meanwhile, CMOB-UB initializes the exploration index $s_\ell = 1$ and the candidate arm set $\mathcal{A}_\ell = [K]$ for each instance $\ell \in [L]$.

With all these preparations, CMOB-UB proceeds to the decision-making process from $t = 1$ to T. In the t-th round, CMOB-UB first samples an instance ℓ_t with probability $2^{-\ell_t}$ for $\ell_t \in [L]$, with the remaining probability sampling $\ell_t = 1$. If the candidate arm set $\mathcal{A}_{\ell_t}$ is not empty, CMOB-UB employs AELO (Algorithm 2) to select an arm a_t. The inputs for AELO are the recorded information of the sampled instance ℓ_t, such as the exploration index s_{ℓ_t}, the candidate arm set $\mathcal{A}_{s_{\ell_t}}$, the empirical rewards $\hat{\mu}^i_{\ell_t}(a)$, and the confidence terms $w_{\ell_t}(a)$ for all $a \in \mathcal{A}_{\ell_t}$.

Once an arm is selected, CMOB-UB eliminates arms of previous instances by intersecting the candidate arm sets $\mathcal{A}_\ell$ and $\mathcal{A}_{\ell_t}$ for all $\ell \leq \ell_t$. This step ensures that the candidate arms from previous instances are contained in subsequent ones, such that $\mathcal{A}_1 \subseteq \mathcal{A}_2 \cdots \subseteq \mathcal{A}_L$. CMOB-UB then plays the chosen arm a_t and receives the corresponding rewards $[\tilde{r}^1_t(a_t), \tilde{r}^2_t(a_t), \ldots, \tilde{r}^m_t(a_t)]$. It updates the empirical mean rewards and the counter as follows:

$$\hat{\mu}^i_{\ell_t}(a_t) = \frac{n_{\ell_t}(a_t) \cdot \hat{\mu}^i_{\ell_t}(a_t) + \tilde{r}^i_t(a_t)}{n_{\ell_t}(a_t) + 1}, n_{\ell_t}(a_t) = n_{\ell_t}(a_t) + 1. \tag{10}$$

Finally, CMOB-UB updates the confidence term $w_{\ell_t}(a_t)$, such that,

$$w_{\ell_t}(a_t) = \sqrt{\frac{\alpha_{\ell_t}(a_t)}{n_{\ell_t}(a_t)}} + \frac{\ln(9MT/\delta)}{n_{\ell_t}(a_t)} \tag{11}$$

where $\alpha_{\ell_t}(a_t) = 4\ln(4MKn_{\ell_t}(a_t)/\delta)$.

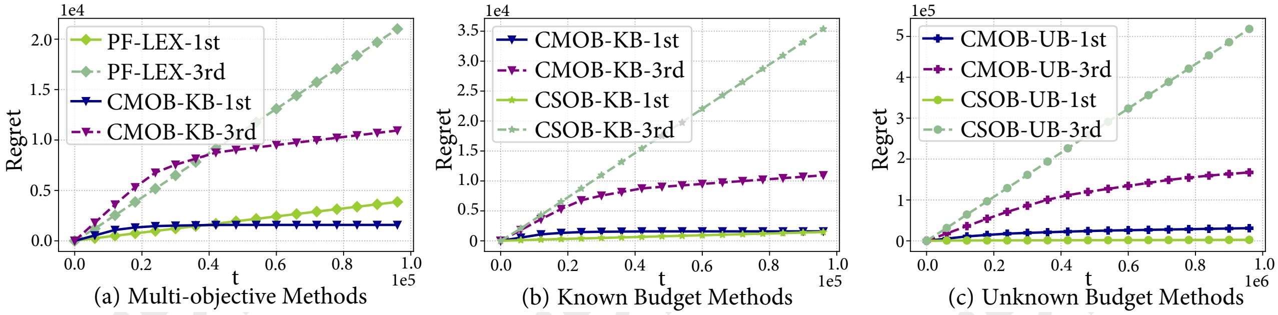
Otherwise, if the candidate arm set $\mathcal{A}_{\ell_t}$ is empty, CMOB-UB finds the minimum instance $\ell'_t \geq \ell_t$ such that $\mathcal{A}_{\ell'_t} \neq \emptyset$ and randomly plays an arm from $\mathcal{A}_{\ell'_t}$. In this case, CMOB-UB does not update the estimated rewards and confidence terms, thus the corruptions does not attack the parameter estimation of CMOB-UB.

We provide the following regret bound for CMOB-UB.

Theorem 2. *Suppose* (3), (4) *and* (5) *hold. If CMOB-UB is run with $\lambda \geq \lambda_0$, then with probability at least $1 - 5\delta$, for any objective $i \in [M]$, its regret satisfies*

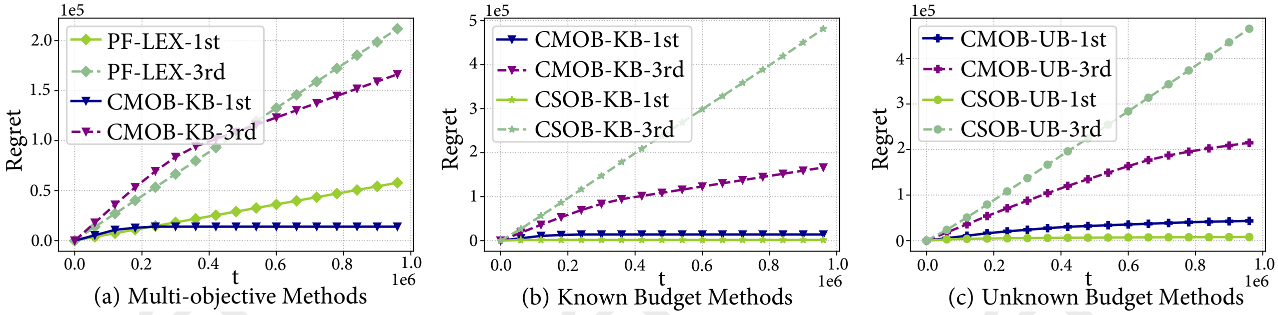
$$R^i(T) \leq \sum_{\Delta^i(a)>0} \alpha_T \beta_T \left(\frac{128(\Lambda^i(\lambda))^2}{\Delta^i(a)} + 16\Lambda^i(\lambda)\right)$$

where $\Lambda^i(\lambda) = 1 + \lambda + \ldots + \lambda^{i-1}$, $\alpha_T = 4\ln(4MKT/\delta)$ and $\beta_T = 4KC\ln(T/\delta) + 2\log_2 T$.



(a) Multi-objective Methods (b) Known Budget Methods (c) Unknown Budget Methods

Figure 2: Comparison of our algorithms versus PF-LEX and single-objective algorithms: Synthetic Dataset and $C = 0$



(a) Multi-objective Methods (b) Known Budget Methods (c) Unknown Budget Methods

Figure 3: Comparison of our algorithms versus PF-LEX and single-objective algorithms: Synthetic Dataset and $C = 300$

Remark 4. Theorem 2 states that CMOB-UB achieves a regret bound of $O\left(\sum_{\Delta^i(a)>0} \gamma_T \left(\frac{(\Lambda^i(\lambda))^2}{\Delta^i(a)} + \Lambda^i(\lambda)\right)\right)$, where $\gamma_T = O\left((KC\log T + \log T)\log(MKT)\right)$. Compared to CMOB-KB, CMOB-UB removes the dependence on the prior knowledge C at the cost of $O(K\log T)$ increase in the regret bound. Compared to the budget-free method for single-objective bandits [Lykouris *et al.*, 2018], CMOB-UB achieves comparable regret bounds in the leading terms $\Delta^i(a)$ and T for multiple objectives simultaneously.

Remark 5. In the case of $C = 0$, the regret bound in Theorem 2 is $O\left(\sum_{\Delta^i(a)>0} (\log(T))^2 \left(\frac{(\Lambda^i(\lambda))^2}{\Delta^i(a)} + \Lambda^i(\lambda)\right)\right)$, which outperforms the existing lexicographic MOMAB regret bound of $\widetilde{O}((KT)^{2/3})$ [Hüyük and Tekin, 2021]. In cases where C is non-zero, the regret bound of CMOB-UB is linear in terms of C, which matches the existing lower bound of corruption-tolerant bandits [Gupta *et al.*, 2019].

6 Experiments

In this section, we conduct experiments on synthetic and real-world datasets to verify the effectiveness of our algorithms.

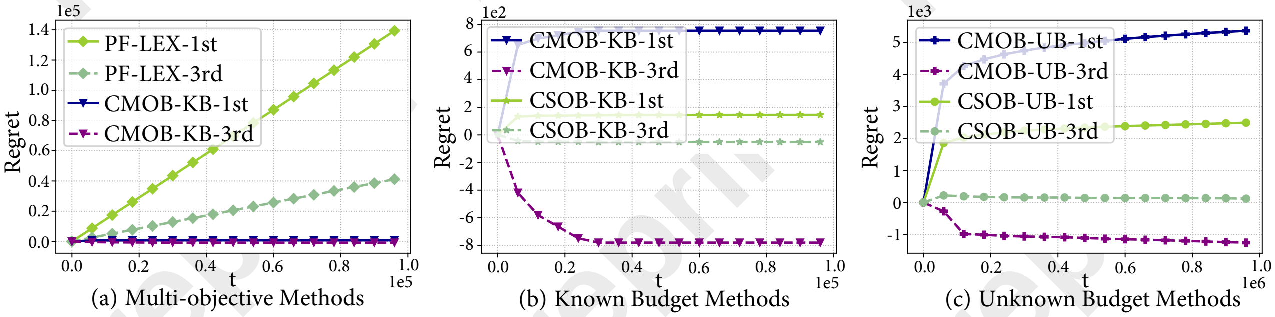
To demonstrate the robustness of our algorithms, we compare them with PF-LEX, a lexicographic MOMAB method for stochastic rewards [Hüyük and Tekin, 2021], and two single-objective corruption-tolerant MAB methods [Lykouris *et al.*, 2018]. Specifically, we refer to the first and third methods from Lykouris *et al.* [Lykouris *et al.*, 2018] as CSOB-KB and CSOB-UB, respectively. CSOB-KB takes the prior knowledge of the budget value C as input, while CSOB-UB removes this dependence.

6.1 Synthetic Dataset

In the synthetic dataset, the number of arms is $K = 20$, and the number of objectives is $M = 3$. We evaluate two levels of corruptions: $C = 0$ for the non-corrupted case and $C = 300$ for the corrupted case. Each algorithm is executed 10 times, and we present the average regrets for the first and third objectives. More details are provided in the appendix.

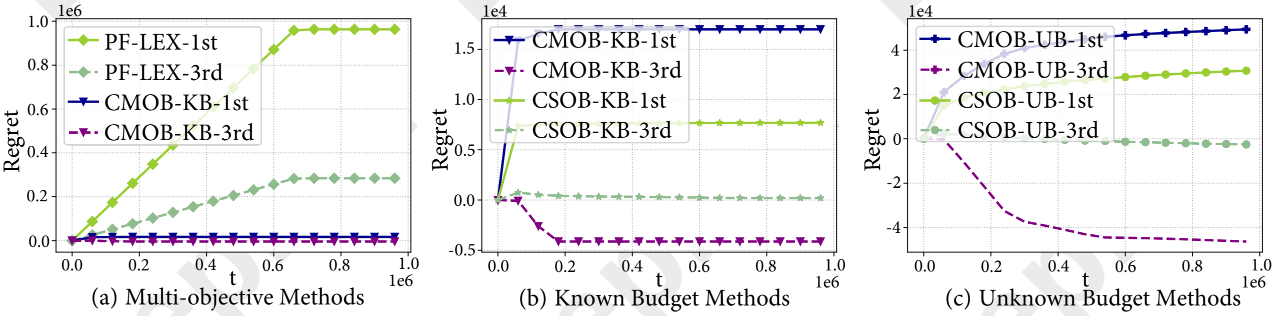
Figure 2 illustrates the non-corrupted case ($C = 0$). In Figure 2(a), we display the performance of PF-LEX and CMOB-KB. It can be observed that CMOB-KB outperforms PF-LEX in both the first and third objectives, which aligns with the theoretical guarantees that CMOB-KB exhibits a lower regret bound than PF-LEX. The regret curve of CMOB-KB eventually flattens, indicating that it successfully identifies the optimal arm. Moving to Figure 2(b), we present the performance of the known budget methods, CMOB-KB and CSOB-KB. Although CMOB-KB and CSOB-KB achieve comparable performance on the first objective, CMOB-KB outperforms CSOB-KB on the third objective, showcasing the effectiveness of CMOB-KB for optimizing multiple objectives. Figure 2(c) presents the results of the unknown budget methods. Similar to Figure 2(b), the multi-objective method CMOB-UB achieves comparable performance to the single-objective algorithm CSOB-UB on the first objective but outperforms CSOB-UB on the third objective.

Figure 3 presents the results for the corrupted case ($C = 300$), with (a), (b), and (c) displaying the multi-objective methods, known budget methods, and unknown budget methods, respectively. The attackers corrupt the rewards and increase the cost of identifying the optimal arm, resulting in the time horizons in Figure 3(a) and Figure 3(b) are 10 times



(a) Multi-objective Methods (b) Known Budget Methods (c) Unknown Budget Methods

Figure 4: Comparison of our algorithms versus PF-LEX and single-objective algorithms: Real-world Dataset and $C = 0$



(a) Multi-objective Methods (b) Known Budget Methods (c) Unknown Budget Methods

Figure 5: Comparison of our algorithms versus PF-LEX and single-objective algorithms: Real-world Dataset and $C = 300$

longer than the zero-corruption case. The curves of CMOB-KB and CMOB-UB eventually converge for both the first and third objectives, showcasing their ability to identify the optimal arm under adversarial attacks and their ability to optimize multiple objectives simultaneously. In contrast, PF-LEX cannot handle corruptions, while CSOB-KB and CSOB-UB yield linear regret bounds for the third objective.

6.2 Real-world Dataset

Building upon the prior research [Kone *et al.*, 2023], we assess the performance of our algorithms using a real-world dataset obtained from Covid-19 vaccines [Munro *et al.*, 2021], where $K = 20$ and $M = 3$. We set the corruption budget C to 0 and 300, respectively. Each algorithm is executed 10 times, and the average regrets are reported.

Figure 4 illustrates the non-corrupted case ($C = 0$). In Figure 4(a), the performance of PF-LEX and CMOB-KB reveals that CMOB-KB outperforms PF-LEX in both the first and third objectives. Figures 4(b) and 4(c) present the performance of methods with known and unknown budgets, respectively. Both of our algorithms, CMOB-KB and CMOB-UB, demonstrate sublinear regret curves in the the first and third objectives, consistent with our theoretical guarantees. Interestingly, unlike in the synthetic experiments, the single-objective algorithms CSOB-KB and CSOB-UB also exhibit sublinear regret curves for the third objective. This is because, in the COV-BOOST dataset, the lexicographic optimal arm can be identified solely based on the first objective. Thus, identifying the optimal arm for the first objective is equivalent to identifying the lexicographic optimal arm in this dataset.

Figure 5 presents the results for the corrupted case ($C = 300$). In comparison to the results in Figure 4, all algorithms exhibit increased regrets due to the attacker corrupts the rewards. Nonetheless, the curves of CMOB-KB and CMOB-UB eventually converge for both the first and third objectives, demonstrating their robustness in identifying the optimal arm.

7 Conclusion and Future Work

We developed two algorithms for lexicographic MAB with adversarial corruptions, enabling the optimization of multiple objectives under attacks. The first algorithm, CMOB-KB, has a regret bound of $O\left(\sum_{\Delta^i(a)>0}\left(\frac{\log(T)(\Lambda^i(\lambda))^2}{\Delta^i(a)}+\Lambda^i(\lambda)C\right)\right)$ for any objective $i \in [M]$, aligning with the single-objective bandit algorithm in terms of $\Delta^i(a), T$ and C [Lykouris *et al.*, 2018]. In the stochastic reward setting, this regret bound reduces to $O\left(\sum_{\Delta^i(a)>0}\frac{\log(T)(\Lambda^i(\lambda))^2}{\Delta^i(a)}\right)$, which improves the existing regret bound $\widetilde{O}((KT)^{2/3})$ [Hüyük and Tekin, 2021]. Although CMOB-KB is simple, its takes the corruption budget C as input. Therefore, we propose the second algorithm, CMOB-UB, which removes the dependence on C and enjoys a regret bound of $O\left(\sum_{\Delta^i(a)>0}\gamma_T\left(\frac{(\Lambda^i(\lambda))^2}{\Delta^i(a)}+\Lambda^i(\lambda)\right)\right)$, where $\gamma_T = O\left(KC(\log T)^2+(\log T)^2\right)$. Finally, we conducted experiments on both synthetic and real-world datasets, verifying the effectiveness of our algorithms in optimizing multiple objectives under corruptions.

One limitation of our algorithms is their reliance on the input value $\lambda \geq \lambda_0$, which may restrict their applicability. Hence, developing an algorithm that does not require prior knowledge of λ_0 would be a valuable advancement.

Acknowledgments

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China [GRF Project No. CityU 11212524].

References

[Altschuler *et al.*, 2019] Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

[Ararat and Tekin, 2023] Cagin Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pages 2165–2190, 2023.

[Auer *et al.*, 2002] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

[Auer *et al.*, 2016] Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 939–947, 2016.

[Auer, 2002] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(11):397–422, 2002.

[Balasubramanian *et al.*, 2024] Rishab Balasubramanian, Jiawei Li, Prasad Tadepalli, Huazheng Wang, Qingyun Wu, and Haoyu Zhao. Adversarial attacks on combinatorial multi-armed bandits. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2505–2526, 2024.

[Bogunovic *et al.*, 2020] Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1071–1081, 2020.

[Bubeck and Cesa-Bianchi, 2012] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[Cai *et al.*, 2023] Xin-Qiang Cai, Pushi Zhang, Li Zhao, Bian Jiang, Masashi Sugiyama, and Ashley J. Llorens. Distributional pareto-optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems 36*, pages 15593–15613, 2023.

[Cheng *et al.*, 2024] Ji Cheng, Bo Xue, Jiaxiang Yi, and Qingfu Zhang. Hierarchize pareto dominance in multi-objective stochastic linear bandits. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 11489–11497, 2024.

[Chowdhury and Gopalan, 2021] Sayak Ray Chowdhury and Aditya Gopalan. No-regret algorithms for multi-task Bayesian optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1873–1881, 2021.

[Deb and Jain, 2014] Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, 18(4):577–601, 2014.

[Ding *et al.*, 2022] Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 7111–7123, 2022.

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

[Ehrgott, 2005] Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2005.

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

[Guan *et al.*, 2020] Ziwei Guan, Kaiyi Ji, Donald J. Bucci Jr., Timothy Y. Hu, Joseph Palombo, Michael Liston, and Yingbin Liang. Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 4036–4043, 2020.

[Gupta *et al.*, 2019] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the 32nd Conference on Learning Theory*, pages 1562–1578, 2019.

[He *et al.*, 2022] Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems 35*, pages 34614–34625, 2022.

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

[Jee *et al.*, 2007] Kyung-Wook Jee, Daniel L. McShan, and Benedick A. Fraass. Lexicographic ordering: intuitive multicriteria optimization for imrt. *Physics in Medicine & Biology*, 52:1845–1861, 2007.

[Kaliszewski, 2000] Ignacy Kaliszewski. Using trade-off information in decision-making algorithms. *Computers & Operations Research*, 27(2):161–182, 2000.

[Kang *et al.*, 2023] Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. In *Advances in Neural Information Processing Systems 36*, pages 10897–10908, 2023.

[Keeney, 2002] Ralph L. Keeney. Common mistakes in making value trade-offs. *Operations Research*, 50(6):935–945, 2002.

[Khansa *et al.*, 2021] Ali Al Khansa, Raphael Visoz, Yezekael Hayel, and Samson Lasaulce. Resource allocation for multi-source multi-relay wireless networks. In *Ubiquitous Networking*, pages 62–75, 2021.

[Kone *et al.*, 2023] Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. In *Advances in Neural Information Processing Systems 36*, pages 35190–35201, 2023.

[Li *et al.*, 2015] Ke Li, Kalyanmoy Deb, Qingfu Zhang, and Sam Kwong. An evolutionary many-objective optimization algorithm based on dominance and decomposition. *IEEE Transactions on Evolutionary Computation*, 19(5):694–716, 2015.

[Lu *et al.*, 2019] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3080–3086, 2019.

[Lu *et al.*, 2021] Shiyin Lu, Guanghui Wang, and Lijun Zhang. Stochastic graphical bandits with adversarial corruptions. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 8749–8757, 2021.

[Lykouris *et al.*, 2018] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, page 114–122, 2018.

[Miettinen, 1999] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, Boston, USA, 1999.

[Mukherjee *et al.*, 2021] Arpan Mukherjee, Ali Tajer, Pin-Yu Chen, and Payel Das. Mean-based best arm identification in stochastic bandits under reward contamination. In *Advances in Neural Information Processing Systems 34*, pages 9651–9662, 2021.

[Munro *et al.*, 2021] Alasdair P S Munro, Leila Janani, Victoria Cornelius, and et al. Safety and immunogenicity of seven covid-19 vaccines as a third dose (booster) following two doses of chadox1 ncov-19 or bnt162b2 in the uk (cov-boost): a blinded, multicentre, randomised, controlled, phase 2 trial. *The Lancet*, 398:2258–2276, 2021.

[Nowak and Trzaskalik, 2022] Maciej Nowak and Tadeusz Trzaskalik. A trade-off multiobjective dynamic programming procedure and its application to project portfolio selection. *Annals of Operations Research*, 311(2):1155–1181, 2022.

[Robbins, 1952] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[Schwartz *et al.*, 2017] Eric Schwartz, Eric Bradlow, and Peter Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(2):500–522, 2017.

[Tekin and Turgay, 2018] Cem Tekin and Eralp Turgay. Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, 66(14):3799–3813, 2018.

[Trisna *et al.*, 2016] Trisna Trisna, Marimin Marimin, Yandra Arkeman, and Titi Sunarti. Multi-objective optimization for supply chain management problem: A literature review. *Decision Science Letters*, 5(6):283–316, 2016.

[Turgay *et al.*, 2018] Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1673–1681, 2018.

[Villar *et al.*, 2015] Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199 – 215, 2015.

[Wang *et al.*, 2023] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[Weber *et al.*, 2002] Enrico Weber, Andrea Emilio Rizzoli, Rodolfo Soncini-Sessa, and Andrea Castelletti. Lexicographic optimisation for water resources planning: the case of lake verbano, italy. pages 235–240, 2002.

[Wilbur and Zhu, 2009] Kenneth Wilbur and Yi Zhu. Click fraud. *Marketing Science*, 28(2):293–308, 2009.

[Xu and Klabjan, 2023] Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *Proceedings of the 40th International Conference on International Conference on Machine Learning*, pages 38499–38517, 2023.

[Xue *et al.*, 2024] Bo Xue, Ji Cheng, Fei Liu, Yimu Wang, and Qingfu Zhang. Multiobjective lipschitz bandits under lexicographic ordering. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 16238–16246, 2024.

[Xue *et al.*, 2025] Bo Xue, Xi Lin, Xiaoyuan Zhang, and Qingfu Zhang. Multiple trade-offs: An improved approach for lexicographic linear bandits. *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 21850–21858, 2025.