# Lexicographic Lipschitz Bandits: New Algorithms and a Lower Bound

**Bo Xue** BOXUE4-C@MY.CITYU.EDU.HK
**Ji Cheng** JCHENG27-C@MY.CITYU.EDU.HK
**Fei Liu** FLIU36-C@MY.CITYU.EDU.HK
*Department of Computer Science, City University of Hong Kong, Hong Kong, China*

**Yimu Wang** YIMU.WANG@UWATERLOO.CA
*Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada*

**Lijun Zhang** ZHANGLJ@LAMDA.NJU.EDU.CN
*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China*
*School of Artificial Intelligence, Nanjing University, Nanjing, China*

**Qingfu Zhang**[*] QINGFU.ZHANG@CITYU.EDU.HK
*Department of Computer Science, City University of Hong Kong, Hong Kong, China*

**Editor:** Tor Lattimore

## Abstract

This paper studies a multiobjective bandit problem under lexicographic ordering, wherein the learner aims to maximize $m$ objectives, each with different levels of importance. First, we introduce the local trade-off, $\lambda_*$, which depicts the trade-off between different objectives. For the case when an upper bound of $\lambda_*$ is known, i.e., $\lambda \geq \lambda_*$, we develop an algorithm that achieves a general regret bound of $\widetilde{O}(\Lambda^i(\lambda)T^{(d_z^i+1)/(d_z^i+2)})$ for the $i$-th objective, where $i \in \{1, 2, \ldots, m\}$, $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$, $d_z^i$ is the zooming dimension for the $i$-th objective, and $T$ is the time horizon. Next, we provide a matching lower bound for the lexicographic Lipschitz bandit problem, proving that our algorithm is *optimal* in terms of $\lambda_*$ and $T$. Finally, for the case where $m = 2$, we remove the dependence on the knowledge about $\lambda_*$, albeit at the cost of increasing the regret bound to $\widetilde{O}(\Lambda^i(\lambda_*)T^{(3d_z^i+4)/(3d_z^i+6)})$, which remains optimal in terms of $\lambda_*$. Compared to existing work on lexicographic multi-armed bandits (Hüyük and Tekin, 2021), our approach improves the current regret bound of $\widetilde{O}(T^{2/3})$ and extends the number of arms to infinity. Numerical experiments confirm the effectiveness of our algorithms.

**Keywords:** Multiobjective Online Learning, Lipschitz Bandits, Lexicographic Order

## 1 Introduction

Online learning with bandit feedback provides a powerful paradigm for modeling sequential decision-making cases (Robbins, 1952), such as clinical trials (Villar et al., 2015), news recommendation (Li et al., 2010), and website optimization (White, 2012). The fundamental model of this paradigm is multi-armed bandits (MAB), where a learner repeatedly selects one arm from $K$ available arms and receives a stochastic payoff drawn from an unknown distribution associated with the chosen arm (Bubeck et al., 2015; Luo et al., 2018; Zhou

[*] Corresponding author.

et al., 2019; Xue et al., 2020; Zhu and Mineiro, 2022; Qin et al., 2023; Gou et al., 2023). The goal of learner is to minimize regret, defined as the cumulative difference between the expected payoff of the selected arm and that of the inherently best arm. To achieve this goal, the learner must strike a balance between exploration and exploitation, attempting potentially better arms while concurrently employing the best arm identified so far.

Although MAB is powerful, many real-world applications involve multiple and potentially conflicting objectives, such as the Click-Through Rate (CTR) and the Post-Click Conversion Rate (CVR) in advertising recommendation systems (Ma et al., 2018). This has led to the study of multiobjective multi-armed bandits (MOMAB), in which payoffs are vectors containing multiple elements, and the learner aims to simultaneously minimize the regret for all objectives (Drugan and Nowe, 2013). A commonly used criterion for evaluating the performance of MOMAB is Pareto regret, which regards all objectives as equivalent (Van Moffaert et al., 2014; Q. Yahyaa et al., 2014; Turgay et al., 2018; Lu et al., 2019b; Xu and Klabjan, 2023). However, some scenarios may require varying levels of importance among objectives, such as radiation treatment for cancer patients, where the primary objective is target coverage and the secondary objective is the therapy's proximity to organs at risk (Jee et al., 2007). Similarly, water resource planning legally mandates the prioritization of objectives such as flood protection, supply shortage for irrigation, and electricity generation (Weber et al., 2002).

To address these real-world applications, a natural idea is to adopt lexicographic ordering, which ranks objectives according to their importance (Ehrgott, 2005; Wray and Zilberstein, 2015; Wray et al., 2015; Hüyük and Tekin, 2021; Hosseini et al., 2021; Skalse et al., 2022; Cheng et al., 2024). Precisely, let $\mathcal{X}$ represent an arm space, and the expected payoffs for $a, b \in \mathcal{X}$ are $[\mu^1(a), \mu^2(a), \dots, \mu^m(a)] \in \mathbb{R}^m$ and $[\mu^1(b), \mu^2(b), \dots, \mu^m(b)] \in \mathbb{R}^m$, respectively. The arm $a$ is said to **lexicographically dominate** arm $b$ if and only if $\mu^1(a) > \mu^1(b)$, or there exists an $i^* \in \{2, 3, \dots, m\}$, such that $\mu^i(a) = \mu^i(b)$ for $1 \leq i \leq i^* - 1$ and $\mu^{i^*}(a) > \mu^{i^*}(b)$. The **lexicographically optimal** arm is the one that is not lexicographically dominated by any other arms (Hüyük and Tekin, 2021).

The only existing work for lexicographic multiobjective bandits is specifically designed for the MOMAB model (Hüyük and Tekin, 2021), whose arm set is finite, i.e., $\mathcal{X} = [K]$[1]. Let $x_*$ denote the lexicographically optimal arm among $\mathcal{X}$ and $x_t$ be the arm chosen at $t$-th epoch. Hüyük and Tekin (2021) defined a priority-based regret to evaluate the performance of their algorithm, given by

$$\widetilde{R}^i(T) = \sum_{t=1}^{T} \left(\mu^i(x_*) - \mu^i(x_t)\right) \mathbb{I}(A^i(x_t)), i \in [m]. \tag{1}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function. $A^i(x_t)$ denotes the event that the previous $i-1$ expected payoffs of the chosen arm are optimal, i.e., $A^i(x_t) = \{\mu^j(x_*) - \mu^j(x_t) = 0, j \in [i-1]\}$. Hüyük and Tekin (2021) proposed an algorithm with a priority-based regret bound of $\widetilde{O}((KT)^{2/3})$.

There are two points for improvement in the existing algorithm (Hüyük and Tekin, 2021). **a)** Its regret bound $\widetilde{O}((KT)^{2/3})$ is suboptimal in its dependence on $T$ when reduced to a single objective, as it fails to match the lower bound $\Omega(K \log T)$ for single objective MAB (Lai and Robbins, 1985). **b)** The regret metric in (1) is inaccurate due to the indicator

[1] For any positive integer $n \in \mathbb{N}_+$, $[n]$ denotes the set $\{1, 2, \dots, n\}$.

function $\mathbb{I}(\cdot)$. Specifically, for the $i$-th objective, if any prior objective $j \in [i-1]$ has suboptimal expected payoff for the chosen arm $x_t$ (i.e., $\mu^j(x_*) \neq \mu^j(x_t)$), the indicator function $\mathbb{I}(A^i(x_t)) = 0$, and the instantaneous regret $\mu^i(x_*) - \mu^i(x_t)$ is not accumulated to the total regret.

Additionally, the MOMAB framework suffers from two inherent constraints: **c)** Its numerical arm representation $(1, 2, \cdots, K)$ cannot incorporate contextual features critical to real-world applications. For example, in news recommendation systems (Li et al., 2010), arms correspond to articles characterized by attributes such as category, length, and author. These features can be effectively incorporated through vector-based arm representations in Lipschitz bandits (Wang et al., 2020; Kang et al., 2023; Feng et al., 2022) or contextual bandits (Xue et al., 2023; Wan et al., 2024; Zhang et al., 2024; Yang et al., 2025), but are disregarded in MOMAB. **d)** Many decision-making problems involve continuous parameters (e.g., hyperparameter tuning (Li et al., 2018), robotic control (Xu et al., 2023)), requiring infinite arm sets. While MOMAB is restricted to finite arms and its regret bound $\widetilde{O}((KT)^{2/3})$ fails for large or infinite $K$, Lipschitz bandits exploit metric space smoothness to generalize insights across arms, enabling infinite arm handling.

To address these issues, we propose a multiobjective Lipschitz bandits (MOLB) framework where the arm set $\mathcal{X}$ is a metric space and the expected payoff functions satisfy the Lipschitz property (Turgay et al., 2018; Wanigasekara and Yu, 2019; Podimata and Slivkins, 2021), such that

$$|\mu^i(x) - \mu^i(x')| \leq \mathcal{D}(x, x'), \forall x, x' \in \mathcal{X}, i \in [m] \tag{2}$$

where $\mathcal{D}(\cdot, \cdot)$ is the distance function on metric space $\mathcal{X}$. Without loss of generality, we assume the diameter of $\mathcal{X}$ is smaller than 1, i.e., $\mathcal{D}(x, x') \leq 1, \forall x, x' \in \mathcal{X}$. To eliminate the indicator function in the metric (1), we introduce a new parameter called the **local trade-off**, defined as

$$\lambda_* = \min \left\{ \lambda \geq 0 \;\middle|\; \mu^i(x) - \mu^i(x_*) \leq \lambda \cdot \max_{j \in [i-1]} \{\mu^j(x_*) - \mu^j(x)\}, \forall i \in [m], \forall x \in \mathcal{X} \right\}. \tag{3}$$

Here, we discuss this parameter from two perspectives. **i) Existence of $\lambda_*$:** In *finite-armed MOMAB*, $\lambda_*$ is finite and well-defined. For each objective $i \in [m]$ and arm $x \in \mathcal{X}$, the inequality $\mu^i(x) - \mu^i(x_*) \leq \lambda \cdot \max_{j \in [i-1]}\{\mu^j(x_*) - \mu^j(x)\}$ imposes an upper bound on $\lambda_*$. Since both the number of objectives $m$ and arms $\mathcal{X}$ are finite in the MOMAB model, the set of such bounds is finite. Consequently, $\lambda_*$ exists as the supremum of these bounds. In *Lipschitz bandits* with infinite arms, $\lambda_*$ may diverge to infinity. Consider a scenario where $\mu(x_*) = [1, 1]$, and for every $n \in \mathbb{Z}_+$, there exists an arm $x_n \in \mathcal{X}$ with $\mu(x_n) = [1 - \frac{1}{n}, 2]$. Here, the trade-off ratio for the second objective becomes

$$\frac{\mu^2(x_n) - \mu^2(x_*)}{\mu^1(x_*) - \mu^1(x_n)} = \frac{2-1}{1 - (1 - \frac{1}{n})} = n.$$

To satisfy Eq. (3), $\lambda$ must exceed $n$ for all $n \in \mathbb{Z}_+$, implying $\lambda_* = +\infty$. **ii) Relationship to Existing Concepts:** $\lambda_*$ aligns conceptually with the established concept called global trade-off in multiobjective optimization (Miettinen, 1999, Definition 2.8.5). In the context of lexicographic bandits, $\lambda_*$ characterizes the trade-offs across the Pareto front. A detailed discussion of the connection between $\lambda_*$ and global trade-offs, $\lambda_*$ and the structure of the

Pareto front, $\lambda_*$ and Lipschitz constant are provided in Section 8. Additionally, Section 3.3 reviews the related work on trade-off concepts in multiobjective optimization.

To the best of our knowledge, this work is the first to explore the MOLB model under lexicographic ordering. We extend the metric of lexicographic bandits from the priority-based regret (1) to the general regret (4), which allows independent evaluation of performance on each objective by removing the indicator function of (1). We propose two parameter-dependent algorithms, SDLO and ADLO, for lexicographic MOLB. These algorithms take a trade-off parameter $\lambda \geq \lambda_*$ as input. When applied to the lexicographic MOMAB problem, SDLO and ADLO achieve regret bounds of $\widetilde{O}(\Lambda^i(\lambda)\sqrt{KT})$ for the $i$-th objective, where $i \in [m]$, $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$. Notably, both SDLO and ADLO improve upon the existing regret bound of $\widetilde{O}((KT)^{2/3})$ in terms of $K$ and $T$ (Hüyük and Tekin, 2021). The key innovation of SDLO and ADLO lies in their multi-stage decision-making approach, which makes a delicate balance between exploration and exploitation. Furthermore, we derive a matching lower bound to demonstrate that our algorithm ADLO achieves the optimal regret bound. Finally, we develop a parameter-free algorithm UCB-ADLO, which removes the requirement about the knowledge of $\lambda_*$ in the case where $m = 2$. UCB-ADLO takes the ADLO algorithm as its base-learner for arm selection, and on top of that it employs a meta-algorithm to select the input trade-off parameter $\lambda$ from a pool of candidates. The main contributions of this work are summarized as follows:

- We adopt the general regret as the metric for lexicographic bandit algorithms, which is
$$R^i(T) = T\mu^i(x_*) - \sum_{t=1}^{T} \mu^i(x_t), i \in [m]. \tag{4}$$
- Equipped with a trade-off parameter $\lambda \geq \lambda_*$, we develop a parameter-dependent algorithm that achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda)T^{(d_z^i+1)/(d_z^i+2)})$ for the $i$-th objective, where $i \in [m]$, $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$, and $d_z^i$ is the zooming dimension of the $i$-th objective (to be defined in the following section).
- We establish a lower bound for the lexicographic MOLB problem, which indicates that our parameter-dependent algorithm is optimal in terms of $\lambda_*$ and $T$.
- For the case $m = 2$, we develop a parameter-free algorithm, which does not require the prior knowledge about $\lambda_*$ but its regret bound is $\widetilde{O}(\Lambda^i(\lambda_*)T^{(3d_z^i+4)/(3d_z^i+6)})$.

An earlier conference version of this paper was presented at the 38th Annual AAAI Conference on Artificial Intelligence in 2024 (Xue et al., 2024). In this extended version, we have enriched the contributions and refined the presentation. First, we provide a formal definition of the local trade-off parameter $\lambda_*$ in Eq. (3). Additionally, we add Section 3.3 to discuss related works on trade-offs in multiobjective optimization and Section 8 to provide further insights on $\lambda_*$. Second, we simplify the filtering operation in Steps 8 and 9 of Algorithms 2 and 4 by replacing the upper confidence bounds with mean payoffs, thus reducing the number of additions. Correspondingly, the theoretical analysis related to the filtering operations has been revised. Third, to demonstrate the importance of the newly proposed parameter $\lambda_*$, we include a matching lower bound for lexicographic MOLB. In addition, we develop a parameter-free algorithm for the case $m = 2$, eliminating the need for prior knowledge about $\lambda_*$. Finally, we conduct additional experiments, including new results for parameter-dependent algorithms with varying input parameters $\lambda \geq \lambda_*$ and for parameter-free algorithms with different local trade-offs $\lambda_*$. The former aims to explore the performance of parameter-dependent algorithms when players have varying degrees of estimation of $\lambda_*$, and the latter examines the performance of the parameter-free algorithm for different lexicographic bandit problems.

## 2 Preliminaries

In this section, we first present the learning setting of MOLB, and then introduce two necessary concepts in Lipschitz bandits, covering dimension and zooming dimension.

### 2.1 Learning Setting

MOLB is a $T$-round decision-making system indexed by $t \in [T]$. At each epoch $t$, the learner selects an arm $x_t$ from the metric space $\mathcal{X}$ and receives a stochastic payoff vector $[y_t^1, y_t^2, \ldots, y_t^m] \in \mathbb{R}^m$, where $y_t^i$ is the payoff of the $i$-th objective and $m$ is the number of objectives. The payoffs are conditionally 1-sub-Gaussian, such that

$$\mathrm{E}\left[e^{\alpha(y_t^i - \mu^i(x_t))} \middle| \mathcal{F}_{t-1}\right] \leq e^{\alpha^2/2}, \forall \alpha \in \mathbb{R} \tag{5}$$

where $\mu^i(x_t)$ denotes the $i$-th expected payoff of arm $x_t$ , $i \in [m]$, and $\mathcal{F}_{t-1} = \{x_1, x_2, \ldots, x_t\} \cup \{y_1^1, y_2^1, \ldots, y_{t-1}^1\} \cup \cdots \cup \{y_1^m, y_2^m, \ldots, y_{t-1}^m\}$ is a $\sigma$-filtration (Auer, 2002; Bubeck et al., 2011b; Abbasi-yadkori et al., 2011; Shao et al., 2018).

### 2.2 Covering Dimension and Zooming Dimension

Let $B(\bar{x}, r)$ denote the ball with center $\bar{x} \in \mathcal{X}$ and radius $r \geq 0$, such that $B(\bar{x}, r) = \{x \in \mathcal{X} | \mathcal{D}(\bar{x}, x) \leq r\}$. The ***r*-covering number** of $\mathcal{X}$ is the minimal number of balls with radius $r$ to cover $\mathcal{X}$, i.e.,

$$N_c(r) = \min\left\{n \in \mathbb{N} \mid \mathcal{X} \subseteq \cup_{k \in [n]} B(\bar{x}_k, r)\right\}. \tag{6}$$

Based on the covering number, the **covering dimension** of $\mathcal{X}$ is defined as

$$d_c = \min\left\{d \geq 0 \;\middle|\; \exists\, C > 0, N_c(r) \leq C r^{-d}, \forall\, r > 0\right\}. \tag{7}$$

We present two specific examples to help with the understanding of covering dimension. One is the unit ball in $d$-dimensional Euclidean space, whose covering dimension is $d$ and $C = 1$. Another example is any set containing finite elements, i.e., $|\mathcal{X}| = K$, whose covering dimension is 0 and $C = K$.

The covering dimension does not account for the structure of expected payoff functions, thus failing to reflect the complexity of a Lipschitz bandit problem accurately. To illustrate this issue, we provide a simple example. Suppose the arm space is $\mathcal{X} \subset \mathbb{R}^d$ with a Euclidean metric, and the expected functions are $\mu^i(x) = x_1, i \in [m]$ for all $x \in \mathcal{X}$. Here, $x_1$ denotes the first element of vector $x$. In this case, no matter how large the covering dimension $d$ is, the complexity of identifying the optimal arm remains the same.

To deal with this issue, another concept termed zooming dimension was proposed (Kleinberg et al., 2008). In this paper, we extend this concept to multiobjective setting. First, we define the ***r*-optimal region** for the $i$-th objective as

$$\mathcal{X}^i(r) = \left\{ x \in \mathcal{X} \;\middle|\; \frac{\Lambda^i(\lambda) r}{2} < \mu^i(x_*) - \mu^i(x) \leq \Lambda^i(\lambda) r \right\} \tag{8}$$

where $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ for a fixed constant $\lambda \geq \lambda_*$ and $r \geq 0$. Then, similar to the $r$-covering number, the ***r*-zooming number** can be defined as the minimal number of balls with radius $r/96$ to cover $\mathcal{X}^i(r)$, denoted by $N_z^i(r)$, i.e.,

$$N_z^i(r) = \min \left\{ n \in N \mid \mathcal{X}^i(r) \subseteq \cup_{k \in [n]} B(\bar{x}_k, r/96) \right\}. \tag{9}$$

Now, we are ready to define the **zooming dimension** for the $i$-th objective, which is

$$d_z^i = \min\{d \geq 0 \mid \exists\, Z_i > 0, N_z^i(r) \leq Z_i r^{-d}, \forall r > 0\}. \tag{10}$$

Compared with the zooming dimension of single objective Lipschitz bandits (Kleinberg et al., 2008), the only difference is the adoption of the constant $\Lambda^i(\lambda)$ in Eq. (8), which is due to technical reasons (see Section 4.2.2) and does not constitute an essential different, as $r$ can approach zero arbitrarily closely.

## 3 Related Work

In this section, we give a brief review of the research for Lipschitz bandits, multiobjective bandits and trade-offs in multiobjective optimization.

### 3.1 Lipschitz Bandits

Plenties of work on Lipschitz bandits have been conducted in recent years, and most of them employ two basic techniques: static discretization (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Magureanu et al., 2014) and adaptive discretization (Kleinberg et al., 2008; Bubeck et al., 2008, 2011a; Lu et al., 2019a; Wang et al., 2020; Feng et al., 2022; Kang et al., 2023). Static discretization involves dividing the arm space into a uniform mesh and directly applying MAB algorithms, such as UCB (Auer, 2002), to the mesh regions. The seminal work of Agrawal (1995) investigated a specific case called continuum-armed bandits, wherein the arm set is a compact interval (i.e., $\mathcal{X} \in [0, 1]$). Building upon this research, Kleinberg (2004) proposed a near-optimal algorithm with a bound of $O(T^{2/3})$ and established a matching lower bound. Subsequently, Auer et al. (2007) improved this result by achieving a regret bound of $O(\sqrt{T})$ under some mild assumptions such that the smoothness of the payoff function is required only at its maxima and that the maxima do not have strong competitors.

Adaptive discretization dynamically discretizes the arm space according to observed payoffs, allocating more trials to promising regions. This technique was first proposed by Kleinberg et al. (2008), who extended the arm set into a general metric space and introduced the zooming algorithm, achieving a regret bound of $\widetilde{O}(T^{(d_z+1)/(d_z+2)})$, where $d_z$ represents the zooming dimension of the expected payoff function. Furthermore, Kleinberg et al. (2008) provided a matching lower bound of $\Omega(T^{(d_z+1)/(d_z+2)})$. A subsequent work of Bubeck et al. (2011a) relaxed the Lipschitz assumption to a locally Lipschitz condition and proposed a tree-based algorithm that attains a regret bound of $\widetilde{O}(T^{(d_z+1)/(d_z+2)})$. Wang et al. (2020) connected tree-based methods with Gaussian processes, developing a new analytical framework. In addition, research has advanced Lipschitz bandits into various directions, such as taxonomy bandits (Slivkins, 2011), contextual Lipschitz bandits (Aleks and rs Slivkins, 2014), Lipschitz bandits with batched feedback (Feng et al., 2022), and Lipschitz bandits with adversarial corruptions (Kang et al., 2023).

### 3.2 Multiobjective Banidts

Drugan and Nowe (2013) initially formalized the MOMAB model and introduced two algorithms enjoying the bounds $O(K \log T)$ under the metrics of scalarized regret and Pareto regret, respectively. Scalarized regret refers to the weighted sum of all objectives' regret, while Pareto regret measures the cumulative Pareto distance between the obtained payoff vectors and the Pareto optimal payoff vector. Turgay et al. (2018) studied the multiobjective contextual bandit model and proposed a zooming-based algorithm that achieves a Pareto regret bound of $\widetilde{O}(T^{(d_p+1)/(d_p+2)})$, where $d_p$ represents the Pareto zooming dimension. Subsequently, Lu et al. (2019b) investigated a parameterized bandit model called multiobjective generalized linear bandits and established a regret bound of $\widetilde{O}(d\sqrt{T})$, where $d$ is the dimension of arm vectors. Another line of research focuses on analyzing the cost of identifying the Pareto optimal arms, known as Pareto Set Identification (PSI) (Auer et al., 2016; Kone et al., 2023, 2024). Auer et al. (2016) introduced key concepts and proposed the first algorithm for PSI. Following this, Kone et al. (2023) addressed different relaxations of the PSI problem and slightly reduced the number of trials needed to identify the Pareto optimal arms. crepon et al. (2024) provided a minimax-type bound on the number of trials required to identify the Pareto optimal arms. Kone et al. (2024) studied the PSI problem under a fixed budget setting, providing an upper bound on the error probability of identifying all Pareto optimal arms. Zhong et al. (2023) designed an adaptive algorithm capable of balancing the trade-off between regret minimization and best arm identification.

In the literature of lexicographic multiobjective bandits, Tekin and Turgay (2018) initially examined contextual bandits with two objectives, achieving a general regret bound of $\widetilde{O}(T^{(d_p+2)/(d_p+3)})$. Later, Hüyük and Tekin (2021) extended the objectives beyond two in the lexicographic MOMAB model, proposing an algorithm enjoying a regret bound of $\widetilde{O}((KT)^{2/3})$. However, this bound is suboptimal as existing single objective MAB algorithms attain a regret bound of $O(K \log T)$ (Lai and Robbins, 1985). Our paper aims to enhance the results of Hüyük and Tekin (2021). To facilitate the understanding of our improvements, we briefly introduce the decision-making strategy of Hüyük and Tekin (2021)'s algorithm, PF-LEX. The fundamental framework to settle the bandit problem is the upper confidence bound (UCB), which first constructs confidence intervals for all arms' expected payoffs, and then selects the arm with the highest upper confidence bound (Lattimore and Szepesvári, 2020). When adapting UCB to the multiobjective bandit context, the main modification is considering all objectives in the arm selection process. Let $c_t(a)$ denote the confidence term of arm $a \in [m]$ at round $t$. PF-LEX considers two cases for arm selection. If some arm $a_t \in [K]$ satisfies $c_t(a_t) > \epsilon$ for a given criterion $\epsilon > 0$, PF-LEX chooses this arm $a_t$. Otherwise, if $c_t(a) < \epsilon$ for all arms $a \in [K]$, PF-LEX filters promising arms based on the confidence intervals sequentially, ranging from the first to the $m$-th objective, and ultimately selects an arm in the $m$-th filtered arm set. PF-LEX consumes numerous trials in the case $c_t(a_t) > \epsilon$, which is a pure exploration scenario and leads to suboptimal regret. Therefore, we propose avoiding the pure exploration case by dividing the decision-making process into multiple stages. Concurrently with our work, the lexicographic bandit problem has also been investigated under adversarial corruption (Xue et al., 2025a) and within the linear model setting (Xue et al., 2025b). In contrast, our work is the first to introduce a parameter-free algorithm and a lower bound for the lexicographic bandit problem.

### 3.3 Trade-offs in Multiobjective Optimization

In the literature on multiobjective optimization, characterizing the trade-offs among objectives is a widely used concept across diverse applications (Athanassopoulos and Podinovski, 1997; Podinovski, 1999; Keeney, 2002; Nowak and Trzaskalik, 2022). The idea of "global trade-off" formalizes this concept by expressing how much one criterion improves when another is reduced by a unit during a transition between all decisions (Miettinen, 1999; Kaliszewski and Michalowski, 1999; Kaliszewski, 2000). A related interpretation is provided by Keeney (2002), who defines "value trade-offs" as the willingness to give up performance on one objective for gains in another, while Ruiz et al. (2019) emphasizes that such trading-off reflects the unavoidable deterioration of one criterion when trying to improve another. These notions have also been applied to the design of decision-support systems (Podinovski, 1999). In our formulation, the trade-off factor $\lambda_*$ plays a similar role: it specifies the amount of improvement in the $i$-th objective that can be obtained per unit sacrifice in the preceding $i-1$ objectives when moving from the lexicographic optimal arm $x_*$ to other sub-optimal arms.

## 4 Parameter-dependent Lexicographic Lipschitz Bandits

In this section, we present two parameter-dependent algorithms for lexicographic Lipschitz bandits along with their theoretical guarantees.

### 4.1 Algorithms

We first introduce an algorithm based on static discretization, which is easy to understand but needs an oracle. Subsequently, we employ adaptive discretization to remove this oracle, thereby devising an algorithm that approaches optimality.

#### 4.1.1 Static Method: SDLO

As a warm-up, we propose a simple algorithm called Static Discretization under Lexicographic Ordering (SDLO), which discretizes the arm set $\mathcal{X}$ before the game starts.

According to the Lipschitz property of expected payoff functions, knowing the expected payoff of $\bar{x} \in \mathcal{X}$ enables us to estimate the expected payoff of any arm $x \in B(\bar{x}, r)$, i.e., $|\mu^i(x) - \mu^i(\bar{x})| \leq r, i \in [m]$. Consequently, a natural strategy for addressing the Lipschitz bandit problem is to discretize the arm space $\mathcal{X}$ into a collection of small balls and identify the best one. Given the radius $r$, using fewer balls to cover the arm space simplifies the task

**Algorithm 1** Static Discretization under Lexicographic Ordering (SDLO)

```
Input: confidence parameter $\delta \in (0,1)$, query radius $r \geq 0$, trade-off parameter $\lambda \geq \lambda_*$
Query the oracle with $r$ to obtain the static arm set $\mathcal{A} = \{\bar{x}_1, \dots, \bar{x}_{N_c(r)}\}$ satisfying $\mathcal{X} \subseteq \cup_{k \in [N_c(r)]} B(\bar{x}_k, r)$
Initialize $\hat{\mu}^i(x) = 0, i \in [m]$ for $x \in \mathcal{A}$
Initialize $r(x) = +\infty$ and $n(x) = 0$ for $x \in \mathcal{A}$
for $t = 1, 2, \dots, T$ do
    Run Algorithm 2 to select an arm $x_t = \text{MSDM-SD}\Big(r, \lambda, \{\hat{\mu}^i(x), r(x)\}_{x \in \mathcal{A}}^{i \in [m]}\Big)$
    Play the arm $x_t$ and receive the payoff $[y_t^1, y_t^2, \dots, y_t^m]$
    Update $\hat{\mu}^i(x_t), i \in [m]$ and $n(x_t)$ according to (14)
    Compute $r(x_t)$ according to (15)
end for
```

of identifying the optimal ball. Thus, covering $\mathcal{X}$ with $N_c(r)$ balls is the best choice, as $N_c(r)$ is the minimum number of balls with radius $r$ needed to cover $\mathcal{X}$. However, constructing this minimal coverage is challenging due to the potentially intricate structure of $\mathcal{X}$. Hence, we assume there exists an oracle that takes radius $r$ as input and outputs the minimal arm set $\mathcal{A} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{N_c(r)}\}$ satisfying

$$\mathcal{X} \subseteq \bigcup_{k \in [N_c(r)]} B(\bar{x}_k, r). \tag{11}$$

Note that for any $x \in \mathcal{X}$, there always exists an arm $\bar{x}_k \in \mathcal{A}$ satisfying $\mathcal{D}(x, \bar{x}_k) \leq r$, which reduces our MOLB problem to a MOMAB problem with $N_c(r)$ arms.

Similar to existing MAB algorithms (Auer, 2002; Yu et al., 2018), SDLO initializes the mean payoffs $\{\hat{\mu}^1(x), \hat{\mu}^2(x), \dots, \hat{\mu}^m(x)\}$ and the counter $n(x)$ to zero for all $x \in \mathcal{A}$, where $n(x)$ counts the times arm $x$ is played. Meanwhile, the confidence term $r(x)$ is initialized to infinity. These terms will be updated with new trial data as learning goes, whose details are given in Eq. (14) and Eq. (15). Equipped with the mean payoffs and confidence terms, SDLO is ready to make a decision. At each epoch $t$, SDLO employs a novel decision-making method to select an arm $x_t$ from $\mathcal{A}$, whose details are outlined in Algorithm 2, referred to as Multi-stage Decision-Making under Static Discretization (MSDM-SD).

Starting with the initialized arm set $\mathcal{A}_1 = \mathcal{A}$ and stage index $s = 1$, MSDM-SD enters a loop that continues until an arm is chosen. In each stage $s$, MSDM-SD first checks if there exists an arm $x_t \in \mathcal{A}_s$ whose confidence term $r(x_t)$ is greater than $2^{-s}$. If such an arm exists, MSDM-SD chooses this arm $x_t$. If no arm in $\mathcal{A}_s$ meets this criterion, MSDM-SD proceeds to an inner loop containing $m$ iterations, which sequentially filters promising arms from the first objective to the $m$-th objective. The initialized arm set for the inner loop is $\mathcal{A}_s^0 = \mathcal{A}_s$. For the $i$-th objective, MSDM-SD first selects the arm $\hat{x}_t^i$ with the highest mean payoff from the previously filtered arm set $\mathcal{A}_s^{i-1}$, i.e.,

$$\hat{x}_t^i = \operatorname*{argmax}_{x \in \mathcal{A}_s^{i-1}} \hat{\mu}^i(x). \tag{12}$$

Then MSDM-SD updates the arm set $\mathcal{A}_s^{i-1}$ to $\mathcal{A}_s^i$ by keeping the promising arms, such that

$$\mathcal{A}_s^i = \left\{x \in \mathcal{A}_s^{i-1} \mid \hat{\mu}^i(x) \geq \hat{\mu}^i(\hat{x}_t^i) - (1 + 2\lambda + \cdots + 2\lambda^{i-1}) \cdot (r + 2 \cdot 2^{-s})\right\}. \tag{13}$$

**Algorithm 2** Multi-stage Decision-Making under Static Discretization (MSDM-SD)

```
Input: query radius $r \geq 0$, trade-off parameter $\lambda \geq \lambda_*$, mean payoffs $\hat{\mu}^i(x)$ and confidence
    terms $r(x)$ for all $i \in [m]$ and $x \in \mathcal{A}$
Initialize $s = 1$ and $\mathcal{A}_1 = \mathcal{A}$
repeat
    if $r(x_t) > 2^{-s}$ for some $x_t \in \mathcal{A}_s$ then
        Choose this arm $x_t$
    else
        Initialize the arm set $\mathcal{A}_s^0 = \mathcal{A}_s$
        for $i = 1, 2, \ldots, m$ do
            $\hat{x}_t^i = \operatorname{argmax}_{x \in \mathcal{A}_s^{i-1}} \hat{\mu}^i(x)$
            $\mathcal{A}_s^i = \left\{x \in \mathcal{A}_s^{i-1} \mid \hat{\mu}^i(x) \geq \hat{\mu}^i(\hat{x}_t^i) - (1 + 2\lambda + \cdots + 2\lambda^{i-1}) \cdot (r + 2 \cdot 2^{-s})\right\}$
        end for
        Update $\mathcal{A}_{s+1} = \mathcal{A}_s^m$ and $s = s + 1$
    end if
until an arm $x_t$ is chosen
Return the chosen arm $x_t$
```

After filtering on the last objective, MSDM-SD sets the arm set $\mathcal{A}_{s+1} = \mathcal{A}_s^m$ and proceeds to the next stage $s = s + 1$. According to Eq. (15), $r(x) > 1/\sqrt{T}$ for all $x \in \mathcal{A}$, MSDM-SD will return an arm $x_t$ to SDLO before $s = \log_2(T)$.

Once SDLO plays the arm $x_t$ and receives payoff vector $[y_t^1, y_t^2, \ldots, y_t^m]$, it updates the mean payoffs $\hat{\mu}^i(x), i \in [m]$ and counter $n(x_t)$ as follows:

$$\hat{\mu}^i(x_t) = \frac{n(x_t)\hat{\mu}^i(x_t) + y_t^i}{n(x_t) + 1}, n(x_t) = n(x_t) + 1. \tag{14}$$

Meanwhile, SDLO updates the confidence term of the chosen arm $x_t$ as

$$r(x_t) = \sqrt{\alpha(x_t)/n(x_t)}. \tag{15}$$

Here, $\alpha(x_t) = 4\ln(4mN_c(r)n(x_t)/\delta)$ and $\delta$ is an input confidence parameter. The following theorem provides a theoretical guarantee for the SDLO algorithm.

**Theorem 1** *Suppose that* (2) *and* (5) *hold. If SDLO is run with* $r \geq 0$ *and* $\lambda \geq \lambda_*$*, then with probability at least* $1 - \delta$*, the regret of SDLO can be bounded as*

$$R^i(T) \leq 2\Lambda^i(\lambda) \cdot \left(rT + 8\sqrt{\alpha_T N_c(r) T}\right), i \in [m]$$

*where* $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ *and* $\alpha_T = 4\ln(mN_c(r)T/\delta)$.

**Remark 1** Theorem 1 states that for any objective $i \in [m]$, SDLO achieves a regret bound of $\widetilde{O}\left(\Lambda^i(\lambda)(rT + \sqrt{N_c(r)T})\right)$. When the arms are finite, i.e., $|\mathcal{X}| = K$, it follows that the query radius $r = 0$ and $N_c(r) = K$. Thus, SDLO attains a regret bound of $\widetilde{O}(\Lambda^i(\lambda)\sqrt{KT})$ for lexicographic MOMAB, which improves the existing regret bound of $\widetilde{O}((KT)^{2/3})$ achieved by PF-LEX (Hüyük and Tekin, 2021). Furthermore, the regret bound of SDLO is based on the general regret Eq. (4), which allows independent evaluation of performance on each objective by removing the indicator function in Eq. (1). Nonetheless, these improvements come with the caveat that SDLO requires the trade-off value $\lambda$ as input, a requirement not needed by PF-LEX.

Recalling the definition of the covering dimension $d_c$ in Eq. (7), $N_c(r) \leq Cr^{-d_c}$ for some constant $C > 0$. By incorporating this inequality into Theorem 1 and minimizing the regret with respect to $r$, we derive a tight bound, as presented below.

**Corollary 1** *Suppose that* (2) *and* (5) *hold. If SDLO is run with* $r = T^{-\frac{1}{2+d_c}}$ *and* $\lambda \geq \lambda_*$, *then with probability at least* $1 - \delta$, *the regret of SDLO can be bounded as*

$$R^i(T) \leq 32\Lambda^i(\lambda) \cdot (\alpha_T C)^{\frac{1}{2}} T^{\frac{1+d_c}{2+d_c}}, i \in [m].$$

Theorem 1 and Corollary 1 reveals that the objective index $i \in [m]$ and parameter $\lambda$ in SDLO exhibit an exponential dependence $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ in the regret bound. This dependence is detrimental for lower-priority objectives when $\lambda > 1$. To mitigate this limitation, we provide an improved regret bound in Appendix A, which reduces sensitivity to the objective index $i$ from exponential to linear under Assumption 1. This improvement is particularly significant for large-scale or high-dimensional optimization tasks where the objective trade-off satisfies Assumption 1. Further details are provided in Appendix A.

### 4.1.2 Adaptive Method: ADLO

Although SDLO is easy to understand, it has two limitations. Firstly, it requires a complicated oracle to discretize the arm space $\mathcal{X}$. Secondly, SDLO fails to match the lower bound of the single objective Lipschitz bandit problem (Kleinberg et al., 2008), indicating room for improvement. Therefore, we adopt the adaptive discretization method proposed by Kleinberg et al. (2008). Our second algorithm, based on adaptive discretization, is called Adaptive Discretization under Lexicographic Ordering (ADLO). The detailed procedure can be found in Algorithm 3.

ADLO maintains an adaptive arm set $\widetilde{\mathcal{A}}$ to construct a collection of balls that cover the arm space $\mathcal{X}$, and the radius of these balls is the confidence term $r(x)$, which is dynamically adjusted as the learning goes. To begin with, ADLO initializes the adaptive arms set $\widetilde{\mathcal{A}}$ with the empty set $\emptyset$. In each round $t$, if the arm space $\mathcal{X}$ is not covered by the set of balls constructed by $\widetilde{\mathcal{A}}$, i.e., $\mathcal{X} \not\subset \cup_{x \in \widetilde{\mathcal{A}}} B(x, r(x))$, ADLO selects an arm $x$ randomly from the uncovered region, adds it to the arm set $\widetilde{\mathcal{A}}$, and plays this arm. The mean payoffs $\{\hat{\mu}^1(x), \hat{\mu}^2(x), \ldots, \hat{\mu}^m(x)\}$ and the counter $n(x)$ of the new arm $x$ are initialized to zero. If the arm space is covered, ADLO employs a multi-stage decision-making method called Multi-Stage Decision-Making under Adaptive Discretization (MSDM-AD) to select the most promising arm from $\widetilde{\mathcal{A}}$.

As shown in Algorithm 4, MSDM-AD takes a similar framework to MSDM-SD, which employs an outer loop to restrict the confidence terms of the arms, and an inner loop to filter promising arms from the first objective to the $m$-th objective. Unlike MSDM-SD, MSDM-AD does not take the query radius $r$ as input, and the candidate arm set $\widetilde{\mathcal{A}}$ changes

**Algorithm 3** Adaptive Discretization under Lexicographic Ordering (ADLO)

```
Input: confidence parameter $\delta \in (0,1)$, trade-off parameter $\lambda \geq \lambda_*$, time horizon $T$
Initialize $\widetilde{\mathcal{A}} = \emptyset$
for $t = 1, 2, \ldots, T$ do
    if $\mathcal{X} \not\subset \cup_{x \in \widetilde{\mathcal{A}}} B(x, r(x))$ then
        Pick an arm $x$ randomly from the uncovered region $\mathcal{X} - \cup_{x \in \widetilde{\mathcal{A}}} B(x, r(x))$
        Update the candidate arm set $\widetilde{\mathcal{A}} = \widetilde{\mathcal{A}} \cup \{x\}$
        Initialize $\hat{\mu}^i(x) = 0, i \in [m]$ and $n(x_t) = 0$
        Play the arm $x_t = x$ and receive the payoff $[y_t^1, \ldots, y_t^m]$
    else
        Run Algorithm 4 to select an arm $x_t = \text{MSDM-AD}\left(\lambda, \{\hat{\mu}^i(x), r(x)\}_{x \in \widetilde{\mathcal{A}}}^{i \in [m]}\right)$
        Play the arm $x_t$ and receive the payoff $[y_t^1, \ldots, y_t^m]$
    end if
    Update $\hat{\mu}^i(x_t), i \in [m]$ and $n(x_t)$ according to (14)
    Compute $r(x_t)$ according to (18)
end for
```

as ADLO goes, resulting in a different filtering mechanism within the inner loop of MSDM-AD. Precisely, for the $i$-th objective, MSDM-AD first selects an arm $\hat{x}_t^i$ that maximizes the mean payoffs, i.e.,

$$\hat{x}_t^i = \underset{x \in \widetilde{\mathcal{A}}_s^{i-1}}{\operatorname{argmax}}\, \hat{\mu}^i(x) \tag{16}$$

where $\widetilde{\mathcal{A}}_s^{i-1}$ is the set filtered on the previous $i-1$ objectives and $\widetilde{\mathcal{A}}_s^0 = \widetilde{\mathcal{A}}_s$. Then, MSDM-AD eliminates arms from $\widetilde{\mathcal{A}}_s^{i-1}$ who are less promising on the $i$-th objective, i.e.,

$$\widetilde{\mathcal{A}}_s^i = \left\{ x \in \widetilde{\mathcal{A}}_s^{i-1} \mid \hat{\mu}^i(x) \geq \hat{\mu}^i(\hat{x}_t^i) - (3 + 6\lambda + \cdots + 6\lambda^{i-1}) \cdot 2^{-s} \right\}. \tag{17}$$

After the inner loop ends, MSDM-AD obtains a set $\widetilde{\mathcal{A}}_s^m$ containing arms that are promising for all $m$ objectives. MSDM-AD then passes $\widetilde{\mathcal{A}}_s^m$ to the next stage $s+1$ as $\widetilde{\mathcal{A}}_{s+1} = \widetilde{\mathcal{A}}_s^m$ for a more refined filtration. Similar to MSDM-SD, MSDM-AD chooses an arm $x_t$ before the stage $s$ increases to $\log(T)$.

Upon playing the arm $x_t$ and receiving the corresponding payoffs, ADLO proceeds to update the mean payoffs $\hat{\mu}^i(x)$ for all $i \in [m]$ according to Eq. (14). Note that the confidence term in SDLO relies on the arm number $N_c(r)$. Since ADLO picks at most $T$ arms, it updates the confidence term as follows,

$$r(x_t) = \sqrt{\tilde{\alpha}(x_t)/n(x_t)} \tag{18}$$

where $\tilde{\alpha}(x_t) = 4\ln(4mTn(x_t)/\delta)$.

There are two main differences between SDLO and ADLO. The first one is how to construct the candidate arm set. SDLO constructs this set statically at the beginning of the algorithm by querying an oracle, while ADLO dynamically expands the arm set as learning progresses. In ADLO, the radius of the circle used to cover the decision set $\mathcal{X}$ shrinks over time, thus more arms need to be introduced. The second difference is the filtering

**Algorithm 4** Multi-stage Decision-Making under Adaptive Discretization (MSDM-AD)

```
Input: trade-off parameter $\lambda \ge \lambda_*$, mean payoffs $\hat{\mu}^i(x)$ and confidence terms $r(x)$ for all
    $i \in [m]$ and $x \in \widetilde{\mathcal{A}}$
Initialize $s = 1$ and $\widetilde{\mathcal{A}}_1 = \widetilde{\mathcal{A}}$
repeat
    if $r(x_t) > 2^{-s}$ for some $x_t \in \widetilde{\mathcal{A}}_s$ then
        Choose this arm $x_t$
    else
        Initialize the arm set $\widetilde{\mathcal{A}}_s^0 = \widetilde{\mathcal{A}}_s$
        for $i = 1, 2, \ldots, m$ do
            $\hat{x}_t^i = \mathrm{argmax}_{x \in \widetilde{\mathcal{A}}_s^{i-1}} \hat{\mu}^i(x)$
            $\widetilde{\mathcal{A}}_s^i = \left\{x \in \widetilde{\mathcal{A}}_s^{i-1} \mid \hat{\mu}^i(x) \ge \hat{\mu}^i(\hat{x}_t^i) - (3 + 6\lambda + \cdots + 6\lambda^{i-1}) \cdot 2^{-s}\right\}$
        end for
        Update $\widetilde{\mathcal{A}}_{s+1} = \widetilde{\mathcal{A}}_s^m$ and $s = s + 1$
    end if
until an arm $x_t$ is chosen
Return the chosen arm $x_t$
```

mechanism in the decision-making stage. MSDM-SD filters arms by the operation (13), which relies on the query radius $r$. In contrast, MSDM-AD employs a filtering mechanism removing the dependence on $r$, as demonstrated by Eq. (17). The following theorem provides a theoretical guarantee for ADLO.

**Theorem 2** *Suppose that* (2) *and* (5) *hold. If ADLO is run with $\lambda \ge \lambda_*$, then with probability at least $1 - \delta$, the regret of ADLO can be bounded as*

$$R^i(T) \le \inf_{r_0 \in (0,1)} \left( \Lambda^i(\lambda) r_0 T + 1152 \Lambda^i(\lambda) \tilde{\alpha}_T \sum_{j \in \mathbb{N}, 2^{-j} \ge r_0} N_z^i(2^{-j}) \cdot 2^j \right), i \in [m]$$

*where $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ and $\tilde{\alpha}_T = 4\ln(4mT^2/\delta)$.*

Recalling the definition of zooming dimension $d_z^i$ in Eq. (10), there exists a constant $Z_i > 0$ such that $N_z^i(r) \le Z_i r^{-d_z^i}$ for any $r > 0$. By applying this inequality to Theorem 2 and minimizing the regret with respect to $r_0 \in (0, 1)$, we derive a tight bound as follows.

**Corollary 2** *Suppose that* (2) *and* (5) *hold. If ADLO is run with $\lambda \ge \lambda_*$, then with probability at least $1 - \delta$, the regret of ADLO can be bounded as*

$$R^i(T) \le 2304 \Lambda^i(\lambda) \tilde{\alpha}_T 2^{d_z^i + 2} Z_i^{\frac{1}{d_z^i + 2}} T^{\frac{d_z^i + 1}{d_z^i + 2}}, i \in [m].$$

**Remark 2** Theorem 2 states that for any objective $i \in [m]$, ADLO attains a regret bound of $\widetilde{O}\left(\Lambda^i(\lambda) T^{(1+d_z^i)/(2+d_z^i)}\right)$. When the arm set is finite, i.e., $|\mathcal{X}| = K$, the zooming dimension $d_z^i = 0$ and the constant $Z_i = K$. In this case, ADLO achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda)\sqrt{KT})$ for the lexicographic MOMAB problem, improving on the existing regret bound of $\widetilde{O}((KT)^{2/3})$ (Hüyük and Tekin, 2021). However, similar to SDLO, ADLO requires access to the trade-off parameter $\lambda$, which limits its applicability compared to PF-LEX. If the input trade-off parameter is minimized as $\lambda_*$, ADLO achieves a regret bound of $\widetilde{O}\left(\Lambda^i(\lambda_*)T^{(1+d_z^i)/(2+d_z^i)}\right)$. In Section 5, we provide a matching lower bound to indicate that ADLO is optimal with respect to $\lambda_*$ and $T$.

Similar to SDLO, ADLO also exhibits an exponential dependence $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ in its regret bound. To address this issue, we derive an improved regret bound in Appendix A, reducing the sensitivity to the objective index $i$ from exponential to linear under Assumption 1. Additional details can be found in Appendix A.

## 4.2 Analysis

In this section, we provide the proof of Theorems 1 and 2. The omitted proof of Corollaries 1 and 2 can be found in the Appendixes B and C, respectively.

### 4.2.1 Proof of Theorem 1

To prove Theorem 1, we begin by presenting three lemmas. The proofs of Lemmas 1 and 3 are detailed in this section, while the proof of Lemma 2 is provided in Appendix D. For clarity, we employ the notations $\hat{\mu}_t^i(x)$, $n_t(x)$ and $r_t(x)$ to represent the value of $\hat{\mu}^i(x)$, $n(x)$ and $r(x)$ at the end of $t$-th round, respectively.

To begin, we present Lemma 1 to show that mean payoff $\hat{\mu}_t^i(x)$ and confidence term $r_t(x)$ construct a reliable confidence interval for the expected payoff $\mu^i(x)$.

**Lemma 1** *With probability at least* $1-\delta$*, for any* $x \in \mathcal{A}$*,*

$$\left|\hat{\mu}_t^i(x) - \mu^i(x)\right| \le r_t(x), \forall i \in [m], t \in [T].$$

**Proof.** The confidence interval of the single-objective MAB (Abbasi-yadkori et al., 2011, Lemma 6) states that for a fixed $i \in [m]$, with probability at least $1-\delta$, for any $x \in \mathcal{A}$,

$$\left|\hat{\mu}_t^i(x) - \mu^i(x)\right| \le c_t(x), t \in [T]$$

where

$$c_t(x) = \sqrt{\left(1 + 2\ln\left(\frac{N_c(r)\sqrt{1+n_t(x)}}{\delta}\right)\right)\frac{1+n_t(x)}{n_t^2(x)}}.$$

By applying a union bound over the $m$ objectives and replacing $\delta$ with $\delta/m$, we have that with probability at least $1-\delta$, for any $x \in \mathcal{A}$,

$$\left|\hat{\mu}_t^i(x) - \mu^i(x)\right| \le \tilde{c}_t(x), \forall i \in [m], t \in [T]$$

where

$$\tilde{c}_t(x) = \sqrt{\left(1 + 2\ln\left(\frac{mN_c(r)\sqrt{1+n_t(x)}}{\delta}\right)\right)\frac{1+n_t(x)}{n_t^2(x)}}.$$

Finally, we employ the fact $1 \leq 2\ln 2$, $\sqrt{1+n_t(x)} \leq 2n_t(x)$ and $1+n_t(x) \leq 2n_t(x)$ to further simplify $\tilde{c}_t(x)$ to $r_t(x)$, which is

$$r_t(x) = \sqrt{4\ln\left(\frac{4mN_c(r)n_t(x)}{\delta}\right)\frac{1}{n_t(x)}}.$$

This concludes the proof of Lemma 1. ■

Next, we introduce Lemma 2 to demonstrate that for any arm $x \in \mathcal{A}_s$, the instantaneous regret decreases exponentially as the stage of MSDM-SD advances. For $i \in [m]$, let $\Delta^i(x) = \mu^i(x_*) - \mu^i(x)$, where $x_*$ is the lexicographically optimal arm.

**Lemma 2** *With probability at least $1-\delta$, for any $x \in \mathcal{A}_s$,*

$$\Delta^i(x) \leq 2\Lambda^i(\lambda)\cdot(r + 2\cdot 2^{-s+1}), i \in [m].$$

Based on Lemma 2, we can further bound the cumulative regret for any arm $x \in \mathcal{A}$ by the following lemma, which illustrates the power of multi-stage decision-making strategy.

**Lemma 3** *With probability at least $1-\delta$, the regret for any $x \in \mathcal{A}$ can be bounded as*

$$n_T(x)\Delta^i(x) \leq 2\Lambda^i(\lambda)\cdot\left(rn_T(x) + 8\sqrt{\alpha_T n_T(x)}\right), i \in [m]$$

*where $\alpha_T = 4\ln(4mN_c(r)T/\delta)$.*

**Proof.** For any $x \in \mathcal{A}$, if $n_T(x) = 1$, Lemma 3 holds trivially. Now, we assume $n_T(x) \geq 2$. Recalling Step 3 of MSDM-SD, if the last time $x$ is chosen occurs at the $s_T(x)$-th stage among the total $T$ rounds, we obtain

$$n_T(x) - 1 \leq 2^{s_T(x)}\sqrt{\alpha_T(n_T(x)-1)}$$

since $x$ is played $n_T(x)-1$ times before this round. Due to the fact that $1 \leq 2^{s_T(x)}\sqrt{\alpha_T n_T(x)}$, we have

$$n_T(x) \leq 2^{s_T(x)+1}\sqrt{\alpha_T n_T(x)}. \tag{19}$$

Multiplying $n_T(x)$ on both sides of the inequality in Lemma 2 yields

$$n_T(x)\Delta^i(x) \leq 2\Lambda^i(\lambda)\cdot\left(rn_T(x) + 2\cdot 2^{-s_T(x)+1}n_T(x)\right). \tag{20}$$

Replacing the second term on the right-hand side of Eq. (20) by Eq. (19) results in

$$n_T(x)\Delta^i(x) \leq 2\Lambda^i(\lambda)\cdot\left(rn_T(x) + 8\sqrt{\alpha_T n_T(x)}\right),$$

which concludes the proof of Lemma 3. ■

Equipped with the Lemmas 1, 2 and 3, we can now complete the proof of Theorem 1. For any objective $i \in [m]$, its regret can be written as

$$R^i(T) = T\mu^i(x_*) - \sum_{t=1}^{T}\mu^i(x_t) = \sum_{x\in\mathcal{A}} n_T(x)\Delta^i(x).$$

Applying Lemma 3 to above equation, we obtain

$$R^i(T) \leq \sum_{x\in\mathcal{A}} 2\Lambda^i(\lambda)\cdot\left(rn_T(x)+8\sqrt{\alpha_T n_T(x)}\right)$$

Due to the fact $\sum_{x\in\mathcal{A}} n_T(x) = T$ and Cauchy-Schwarz inequality[2], we obtain that

$$R^i(T) \leq 2\Lambda^i(\lambda)\cdot\left(rT+8\sqrt{\alpha_T|\mathcal{A}|T}\right) = 2\Lambda^i(\lambda)\cdot\left(rT+8\sqrt{\alpha_T N_c(r)T}\right).$$

This concludes the proof of Theorem 1. ■

### 4.2.2 Proof of Theorem 2

Similar to the proof of Theorem 1, we use the notation $\hat{\mu}_t^i(x)$, $n_t(x)$, and $r_t(x)$ to represent the values of $\hat{\mu}^i(x)$, $n(x)$, and $r(x)$ at the end of the $t$-th epoch, respectively. Furthermore, $\widetilde{\mathcal{A}}_t$ denotes the adaptive arm set $\widetilde{\mathcal{A}}$ at the end of the $t$-th epoch, and $\widetilde{\mathcal{A}}_{t,s}$ represents the arm set $\widetilde{\mathcal{A}}_s$ in MSDM-AD. To prove Theorem 2, we need to conduct four lemmas first. The omitted proof of Lemmas 4, 5 and 6 are provided in Appendixes E, F and G, respectively.

To begin, we show that the mean payoffs and confidence terms in ADLO are confident for the inherent expected payoffs.

**Lemma 4** *With probability at least* $1-\delta$*, for any* $x\in\widetilde{\mathcal{A}}_t$*,*

$$\left|\hat{\mu}_t^i(x)-\mu^i(x)\right| \leq r_t(x), i\in[m], t\in[T].$$

Next, we demonstrate an essential property of the multi-stage decision-making strategy.

**Lemma 5** *With probability at least* $1-\delta$*, for any* $s\geq 1$ *and* $x\in\widetilde{\mathcal{A}}_{t,s}$*,*

$$\Delta^i(x) \leq 6\Lambda^i(\lambda)\cdot 2^{-s+1}, i\in[m], t\in[T].$$

For any objective $i\in[m]$, the suboptimal arms constitute the set $\widetilde{\mathcal{A}}_+^i = \{x\in\widetilde{\mathcal{A}}_T \mid \Delta^i(x) > 0\}$. To proceed with the analysis, we partition the suboptimal arm set $\widetilde{\mathcal{A}}_+^i$ into a set of disjoint subsets. Specifically, we define

$$\widetilde{\mathcal{A}}_j^i = \left\{x\in\widetilde{\mathcal{A}}_+^i \mid \Lambda^i(\lambda)2^{-j-1} < \Delta^i(x) \leq \Lambda^i(\lambda)2^{-j}\right\}, \tag{21}$$

thus $\widetilde{\mathcal{A}}_+^i = \cup_{j\in\mathbb{N}}\widetilde{\mathcal{A}}_j^i$. Recall the definitions of $r$-optimal region in Eq. (8) and zooming dimension $d_z^i$ in Eq. (10), we can bound the number of arms in $\widetilde{\mathcal{A}}_j^i$ by the following lemma.

**Lemma 6** *With probability at least* $1-\delta$*, for any* $j\in\mathbb{N}$*,*

$$|\widetilde{\mathcal{A}}_j^i| \leq N_z^i(2^{-j}), i\in[m].$$

In the following, we give Lemma 7 to bound the cumulative regret of any arm in $\widetilde{\mathcal{A}}_j^i$. The proof of Lemma 7 illustrates the advantage of dividing the decision-making process into multiple stages.

[2] For any sequences $\{a_i\}_{i\in[n]}\subseteq\mathbb{R}$ and $\{b_i\}_{i\in[n]}\subseteq\mathbb{R}$, $\sum_{i=1}^n a_ib_i \leq \sqrt{\sum_{i=1}^n a_i^2}\cdot\sqrt{\sum_{i=1}^n b_i^2}$.

**Lemma 7** *With probability at least $1-\delta$, for all $j \in \mathbb{N}$, the regret for any $x \in \widetilde{\mathcal{A}}_j^i$ can be bounded as*

$$n_T(x)\Delta^i(x) \leq 1152\Lambda^i(\lambda)\tilde{\alpha}_T \cdot 2^j, i \in [m]$$

*where $\tilde{\alpha}_T = 4\ln(4mT^2/\delta)$.*

**Proof.** For any $x \in \widetilde{\mathcal{A}}_j^i$, if $n_T(x) = 1$, Lemma 7 holds trivially. Now, we assume $n_T(x) \geq 2$. Let $s_T(x)$ be the stage index that the last time $x$ is chosen. Recalling Step 3 of MSDM-AD, since $x$ is played $n_T(x) - 1$ times before this round, we get

$$n_T(x) - 1 \leq 2^{s_T(x)}\sqrt{2\tilde{\alpha}_T(n_T(x) - 1)}$$

Then, due to the fact that $1 \leq 2^{s_T(x)}\sqrt{\tilde{\alpha}_T n_T(x)}$, we have

$$n_T(x)\Delta^i(x) \leq 2^{s_T(x)+1}\sqrt{\tilde{\alpha}_T n_T(x)}\Delta^i(x). \tag{22}$$

Taking Lemma 5 into the right-hand side of (22) yields

$$n_T(x)\Delta^i(x) \leq 24\Lambda^i(\lambda)\sqrt{\tilde{\alpha}_T n_T(x)}. \tag{23}$$

This step reduces the linear term $n_T(x)\Delta^i(x)$ to a sublinear term $\widetilde{O}(\sqrt{n_T(x)})$, which serves as the crucial function for dividing the decision-making process into multiple stages. Squaring both sides of (23) gives

$$n_T(x)\Delta^i(x) \leq 576(\Lambda^i(\lambda))^2\tilde{\alpha}_T/\Delta^i(x).$$

The definition of $\widetilde{\mathcal{A}}_j^i$ implies that $1/\Delta^i(x) < 2^{j+1}/\Lambda^i(\lambda)$ for any $x \in \widetilde{\mathcal{A}}_j^i$. Taking it into the right-hand side of the above equation finishes the proof of Lemma 7. ■

Now, we are ready to prove Theorem 2. First, we relax $R^i(T)$ by some $r_0 > 0$ as follows,

$$R^i(T) \leq \Lambda^i(\lambda)r_0T + \sum_{x\in\widetilde{\mathcal{A}}_+^i} n_T(x)\Delta^i(x) \cdot \mathbb{I}(\Delta^i(x) > \Lambda^i(\lambda)r_0).$$

According to $\widetilde{\mathcal{A}}_+^i = \cup_{j\in\mathbb{N}}\widetilde{\mathcal{A}}_j^i$ and the definition of $\widetilde{\mathcal{A}}_j^i$ in Eq. (21), we rewrite the above equation as

$$R^i(T) \leq \Lambda^i(\lambda)r_0T + \sum_{j\in\mathbb{N},2^{-j}\geq r_0}\sum_{x\in\widetilde{\mathcal{A}}_j^i} n_T(x)\Delta^i(x).$$

Based on Lemma 6 and Lemma 7, we can easily obtain

$$R^i(T) \leq \Lambda^i(\lambda)r_0T + 1152\Lambda^i(\lambda)\tilde{\alpha}_T \sum_{j\in\mathbb{N},2^{-j}\geq r_0} N_z^i(2^{-j}) \cdot 2^j.$$

This concludes the proof of Theorem 2. ■

## 5 Lower Bound

In this section, we provide a lower bound for the lexicographic Lipschitz bandit problem, which indicates our algorithm ADLO is optimal in terms of $\lambda_*$ and $T$. First, we introduce two key quantities. The first one is

$$R_c(T) = \inf_{r_0 \in (0,1)} \left( r_0 T + \ln T \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} N_c(2^{-j}) \cdot 2^j \right) \tag{24}$$

where $N_c(2^{-j})$ is the $2^{-j}$-covering number of $\mathcal{X}$, defined precisely in Eq. (6). The second one is

$$R_z^i(T) = \inf_{r_0 \in (0,1)} \left( \Lambda^i(\lambda_*) r_0 T + \Lambda^i(\lambda_*) \ln T \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} N_z^i(2^{-j}) \cdot 2^j \right)$$

where $i \in [m]$ and $N_z^i(2^{-j})$ is the $2^{-j}$-zooming number for the $i$-th objective, defined precisely in Eq. (9).

We formalize the notion of $r$-packing: A subset $S \subseteq \mathcal{X}$ is an $r$**-packing** of $\mathcal{X}$ if the distance between any two points in $S$ is at least $r$, i.e., $\inf_{u,v \in S} \mathcal{D}(u,v) \geq r$. The $r$**-packing number** $N_p(r)$ is the maximum cardinality of such subsets:

$$N_p(r) = \max \left\{ |S| \;\middle|\; S \subseteq \mathcal{X}, \inf_{u,v \in S} \mathcal{D}(u,v) \geq r \right\}.$$

**Theorem 3** *For any bandit algorithm $\mathcal{B}$, there exists at least one problem instance on which the expected regret of $\mathcal{B}$ satisfies*

$$\mathrm{E}[R^i(T)] \geq \frac{R_z^i(T)}{320}, i \in [m].$$

**Proof.** Following established lower bound techniques (Bubeck and Cesa-Bianchi, 2012; Aleks and rs Slivkins, 2014; Lu et al., 2019a), we construct a set of problem instances with distinct lexicographic optimal arms that are statistically indistinguishable. Consequently, any algorithm $\mathcal{B}$ will inevitably select suboptimal arms in at least one instance, thereby establishing a lower regret bound.

We select a constant $R \leq R_c(T)$. Let $\tilde{r} = \frac{R}{6T \ln T}$ and $N = \max\left\{2, \lfloor T \cdot \tilde{r}^2 \rfloor\right\}$. The following lemma is proposed to construct the problem instances.

**Lemma 8** *If $T > 2$ and $R \leq R_c(T)$, then $\tilde{r} \leq 1/2$ and $N \leq N_p(\tilde{r})$.*

Based on Lemma 8, we can find a set of arms $U = \{u_1, \ldots, u_N\} \subseteq \mathcal{X}$ such that $\inf_{x,y \in U} \mathcal{D}(x,y) \geq \tilde{r}$. Next, we construct the set of problem instances $\mathcal{I}$. For a given problem instance $I_n \in \mathcal{I}$, the expected payoff function of the $i$-th objective is specified as follows:[3]

[3] The 1-Lipschitz continuity of Eq. (25) is proved in Appendix K.

$$\mu_n^i(x) = \begin{cases} \frac{7}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}, & x = u_n \\ \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}, & x = u_j, j \in [N], j \neq n. \\ \max\left\{ \frac{1}{2} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}, \max_{u \in U} \mu_n^i(u) - \mathcal{D}(x,u) \right\}, & \text{otherwise} \end{cases} \tag{25}$$

The associated payoff distributions are defined by

$$\Pr(y|x) = p_n^i(y|x) = \begin{cases} \frac{\mu_n^i(x)}{\Lambda^i(\lambda_*)}\tilde{r}, & y = \frac{\Lambda^i(\lambda_*)}{\tilde{r}} \\ 1 - \frac{\mu_n^i(x)}{\Lambda^i(\lambda_*)}\tilde{r}, & y = 0 \end{cases}. \tag{26}$$

Recall the definition of the $r$-optimal region $\mathcal{X}^i(r) = \left\{ x \in \mathcal{X} \mid \frac{\Lambda^i(\lambda_*)r}{2} < \Delta^i(x) \leq \Lambda^i(\lambda_*)r \right\}$. For any instance in $\mathcal{I}$ and $r \geq \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{1}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$, it follows that $\mathcal{X}^i(r) = \emptyset$ and $N_z^i(r) = 0$. Recall that $\tilde{r} = \frac{R}{6T \ln T}$, we have

$$\begin{aligned} R_z^i(T) &\leq \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}T + \Lambda^i(\lambda_*) \ln T \sum_{j \in \mathbb{N}, 2^{-j} \geq \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{1}{\Lambda^m(\lambda_*)} \cdot \tilde{r}} N_z^i(2^{-j}) \cdot 2^j \\ &\leq \frac{1}{12m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*) R}{\Lambda^m(\lambda_*) \ln T}. \end{aligned} \tag{27}$$

Next, we construct a special problem instance $I_0$ which is similar to the problem instances in $\mathcal{I}$, but with different optimal arms. In instance $I_0$, the expected payoff function of the $i$-th objective is defined as:

$$\mu_0^i(x) = \begin{cases} \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}, & x = u_j, j \in [N] \\ \max\left( \frac{1}{2} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}, \max_{u \in U} \mu_0^i(u) - \mathcal{D}(x,u) \right), & \text{otherwise} \end{cases}. \tag{28}$$

The associated payoff distributions are defined by

$$\Pr(y|x) = p_0^i(y|x) = \begin{cases} \frac{\mu_0^i(x)}{\Lambda^i(\lambda_*)}\tilde{r}, & y = \frac{\Lambda^i(\lambda_*)}{\tilde{r}} \\ 1 - \frac{\mu_0^i(x)}{\Lambda^i(\lambda_*)}\tilde{r}, & y = 0 \end{cases}. \tag{29}$$

Let the sample space be defined as $\Omega = \mathcal{X} \times \{0, \Lambda^1(\lambda_*)/\tilde{r}\} \times \{0, \Lambda^2(\lambda_*)/\tilde{r}\} \times \cdots \times \{0, \Lambda^m(\lambda_*)/\tilde{r}\}$. For notational clarity, we denote $\Omega^t$ as the $t$-fold product space of $\Omega$ for any $t \geq 1$. The payoff vector received at the $t$-th round is denoted by $\boldsymbol{y}_t = [y_t^1, y_t^2, \ldots, y_t^m]$. For any algorithm $\mathcal{B}$ executed on problem $I_n$, the history trials up to round $t$ are represented by $h_t = [(x_1, \boldsymbol{y}_1), \ldots, (x_t, \boldsymbol{y}_t)] \in \Omega^t$, which is a realization of a random variable defined on the probability space $(\Omega^t, \mathcal{E}, Q_t^n)$. Here, $\mathcal{E}$ denotes the event space generated by the sample space $\Omega^t$, and $Q_t^n$ is the probability measure induced by algorithm $\mathcal{B}$ and problem $I_n$.

For $n = 1, 2, \ldots, N$, define the set $S_n = B(u_n, 3\tilde{r}/8)$. Since $\mathcal{E}$ is the event space induced by $\Omega^t$, the event $\{x_t \in S_n\}$ belongs to $\mathcal{E}$. We define the indicator random variable $X_n(t) = \mathbb{I}(x_t \in S_n)$, which is measurable by construction. Let $Z_n = \sum_{t=1}^{T} X_n(t)$. Since $S_1, S_2, \ldots, S_N$ are mutually disjoint, it follows that $\sum_{n=1}^{N} \mathrm{E}_{Q_T^0}[Z_n] \leq T$. Therefore, there must exist $\tilde{n} \in [N]$ such that

$$\mathrm{E}_{Q_T^0}[Z_{\tilde{n}}] \leq T/N. \tag{30}$$

Due to the special design of the problems (26) and (29), the similarity between instances $I_0$ and $I_{\tilde{n}}$ can be quantified by the following lemma.

**Lemma 9** *The Kullback–Leibler divergence from $Q_T^0$ to $Q_T^{\tilde{n}}$ satisfies $KL(Q_T^0, Q_T^{\tilde{n}}) \leq 39/200$.*

With Lemma 9 in hand, we are ready to finish the proof of Theorem 3. Let $B$ denote the event that $Z_{\tilde{n}} \leq \frac{7T}{4N}$ and $^\neg B$ be the event that $Z_{\tilde{n}} > \frac{7T}{4N}$. By the Markov's inequality (Huber, 2019), we have

$$Q_T^0\left(^\neg B\right) \leq \frac{\mathrm{E}_{Q_T^0}[Z_{\tilde{n}}]}{7T/(4N)} \leq \frac{T/N}{7T/(4N)} = \frac{4}{7}.$$

Therefore, $Q_T^0\left(B\right) \geq 3/7$. By the Pinsker's inequality (Tsybakov, 2008), we obtain that

$$Q_T^{\tilde{n}}\left(B\right) \geq Q_T^0\left(B\right) - \sqrt{KL(Q_T^0, Q_T^{\tilde{n}})/2} \geq 1/10,$$

which implies

$$\mathrm{E}_{Q_T^{\tilde{n}}}[T - Z_{\tilde{n}}] \geq Q_T^{\tilde{n}}\left(B\right) \mathrm{E}_{Q_T^{\tilde{n}}}[T - Z_{\tilde{n}} | B] \geq \frac{1}{10}\left(T - \frac{7T}{4N}\right) \geq \frac{T}{80}.$$

For the problem $I_{\tilde{n}}$, whenever the algorithm $\mathcal{B}$ plays an arm that is not in $S_{\tilde{n}}$, the $i$-th objective suffers an instantaneous regret bigger than $\frac{1}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$. Hence, we have

$$\begin{aligned}
\mathrm{E}_{Q_T^{\tilde{n}}}[R^i(T)] &\geq \frac{1}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r} \cdot \mathrm{E}_{Q_T^{\tilde{n}}}[T - Z_{\tilde{n}}] \\
&\geq \frac{1}{640} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}T \\
&= \frac{1}{3840} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \frac{R}{\ln T}.
\end{aligned} \tag{31}$$

Combining Eq. (31) and Eq. (27), we get

$$\mathrm{E}_{Q_T^{\tilde{n}}}[R^i(T)] \geq \frac{R_z^i(T)}{320}.$$

The proof of Theorem 3 is finished. ■

**Algorithm 5** UCB-ADLO

```
Input: confidence parameter $\delta \in (0,1)$, zooming parameters $d_z^1$ and $Z_1$, time horizon $T$
Initialize $H = \lfloor T^{2/3} \rfloor$, $\widetilde{K} = \lceil \log_2 H \rceil + 1$
Set $\lambda_k = 2^{k-1}$ for $k \in [\widetilde{K}]$
for $k = 1, 2, \ldots, \widetilde{K}$ do
    Restart Algorithm 3 for $H$ rounds with confidence parameter $1/H$ and trade-off parameter $\lambda_k$ and receive the payoffs $[y_t^1, y_t^2], t = (k-1)H+1, (k-1)H+2, \ldots, kH$
    Accumulate payoffs: $Y^i(k) = \sum_{t=(k-1)H+1}^{kH} y_t^i$ for any $i = 1, 2$
end for
$\mathcal{K}^1 = \left\{ k \in [\widetilde{K}] \middle| Y^1(\widetilde{K}) - Y^1(k) \le 4\sqrt{\ln(8\widetilde{K}/\delta)} + 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}} \right\}$
$k^\dagger = \mathrm{argmax}_{k \in \mathcal{K}^1} Y^2(k)$
for $\tau = \widetilde{K}, \widetilde{K}+1, \ldots, \lceil T/H \rceil$ do
    Restart Algorithm 3 for $H$ rounds with confidence parameter $1/H$ and trade-off parameter $\lambda_{k^\dagger}$
end for
```

## 6 Parameter-free Lexicographic Lipschitz Bandits

Previously, we introduced two lexicographic Lipschitz bandit algorithms, SDLO and ADLO, both of which require to know an upper bound of the local trade-off parameter $\lambda_*$. This requirement may limit their applicability when such prior knowledge is unavailable. In this section, we propose a parameter-free lexicographic Lipschitz bandit algorithm that removes the need for knowledge of $\lambda_*$ in the two-objective case ($m = 2$). We also derive a corresponding regret bound for this new algorithm.

### 6.1 Algorithm

To eliminate the dependence on $\lambda_*$, we adopt a meta-learning framework widely used in dynamic online learning (Daniely et al., 2015; Jun et al., 2017b; Zhang et al., 2018; Cheung et al., 2019; Zhao et al., 2021). Inspired by the Bandits-over-Bandits (BOB) mechanism (Cheung et al., 2019), we design a UCB-based meta-algorithm that learns $\lambda_*$ adaptively. This approach partitions the $T$-round learning process into multiple subprocesses, where each subprocess selects a trade-off parameter $\lambda$ using the UCB strategy (Auer et al., 2002).

Specifically, we take the ADLO algorithm as the base learner for arm selection, while on top of that a meta-algorithm dynamically adjusts the input parameter $\lambda$. This parameter-free algorithm is called UCB-ADLO, and its details are shown in Algorithm 5.

Initially, UCB-ADLO sets the sub-horizon $H = \lfloor T^{2/3} \rfloor$ and the number of trade-off parameters $\widetilde{K} = \lceil \log_2 H \rceil + 1$. The candidate trade-off parameters are then defined as

$$\lambda_k = 2^{k-1}, \quad k \in [\widetilde{K}].$$

UCB-ADLO treats these candidate parameters as arms in bandit framework, where executing ADLO with $\lambda_k$ corresponds to "playing arm $k \in [\widetilde{K}]$". Precisely, for each $k \in [\widetilde{K}]$, UCB-ADLO runs ADLO for $H$ rounds using $\lambda_k$ and computes cumulative payoffs:

$$Y^i(k) = \sum_{t=(k-1)H+1}^{k\times H} y_t^i, \quad i \in [m].$$

Since ADLO is a deterministic algorithm, under the trade-off parameter $\lambda_k$, the expected payoff for each objective $i \in [m]$ lies within the following confidence interval with probability at least $1-\delta$:

$$\left[ Y^i(k) - 2\sqrt{\ln(8\widetilde{K}/\delta)}, Y^i(k) + 2\sqrt{\ln(8\widetilde{K}/\delta)} \right]. \tag{32}$$

By construction, $\lambda_{\widetilde{K}} \geq \lambda_*$. Otherwise, $H \leq \lambda_{\widetilde{K}} < \lambda_*$ would lead to linear regret in the second objective, rendering the multi-objective setting meaningless. Thus, Theorem 2 ensures that running ADLO with $\lambda_{\widetilde{K}}$ achieves the optimal regret bound for the first objective.

Using the confidence interval in Eq. (32), UCB-ADLO filters promising trade-off parameters based on their performance on the first objective, such as

$$\mathcal{K}^1 = \left\{ k \in [\widetilde{K}] \middle| Y^1(\widetilde{K}) - Y^1(k) \leq 4\sqrt{\ln(8\widetilde{K}/\delta)} + 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{-1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}} \right\}.$$

Lemma 11 ensures that $\mathcal{K}^1$ contains the optimal parameter $k'$ approximating the local trade-off $\lambda_*$, such that $\lambda_{k'}/2 < \lambda_* \leq \lambda_{k'}$.

Based on this, UCB-ADLO then selects the final trade-off parameter from $\mathcal{K}^1$ by maximizing the payoff of the second objective:

$$k^\dagger = \operatorname*{argmax}_{k\in\mathcal{K}^1} Y^2(k).$$

Since $k' \in \mathcal{K}^1$, the performance of $\lambda_{k^\dagger}$ on the second objective is no worse than that of $\lambda_{k'}$. Given that $\lambda_{k'}$ optimally approximates $\lambda_*$, the performance of $\lambda_{k^\dagger}$ remains comparable to $\lambda_*$. Finally, in each subsequent iteration $\tau = \widetilde{K}+1$ to $\lceil T/H \rceil$, UCB-ADLO executes ADLO with$\lambda_{k^\dagger}$ for $H$ rounds.

The theoretical guarantees for UCB-ADLO are formalized in the following theorem.

**Theorem 4** *Suppose that* (2) *and* (5) *hold. If the objective number $m = 2$ and $\lambda_* \geq 1$, with probability at least $1-\delta$, for any objective $i \in [m]$, its expected regret of UCB-ADLO can be bounded as*

$$\mathrm{E}[R^i(T)] \leq \widetilde{K}T^{\frac{2}{3}} + 16\sqrt{\ln(8\widetilde{K}/\delta)} \cdot T^{\frac{1}{3}} + 36864\Lambda^i(\lambda_*)\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{-1}{d_z^i+2}} T^{\frac{3d_z^i+4}{3d_z^i+6}}$$

*where $\widetilde{K} = \lceil \log_2 H \rceil + 1$, $\Lambda^i(\lambda_*) = 1 + \lambda_* + \cdots + \lambda_*^{i-1}$ and $\bar{\alpha}_H = 12\ln(2H)$.*

**Remark 3** Theorem 4 states that if $m = 2$ and $\lambda_* \geq 1$, UCB-ADLO attains a regret bound of $\widetilde{O}(\Lambda^i(\lambda_*)T^{(3d_z^i+4)/(3d_z^i+6)})$ for the $i$-th objective. This bound aligns with our proposed lower bound $\Omega(R_z^i(T))$ in terms of $\lambda_*$, while requiring no prior knowledge about $\lambda_*$.

**Remark 4** If $\lambda_* < 1$, then $\lambda_k \geq \lambda_*$ for all $k \in [\widetilde{K}]$. Therefore, we have that $\mathcal{K}^1 = [\widetilde{K}]$ using a similar argument as in Lemma 11. This implies that the regret of the second objective is $\widetilde{O}(\Lambda^2(\lambda_1)T^{(3d_z^2+4)/(3d_z^2+6)})$. Since Theorem 3 gives a lower bound of $\Omega(\Lambda^2(\lambda_*))$, the gap for the second objective is $O(1-\lambda_*)$, as $\Lambda^2(\lambda_1) - \Lambda^2(\lambda_*) = 1 - \lambda_*$.

**Remark 5** If the number of arms is finite, i.e., $|\mathcal{X}| = K$, then the zooming dimension $d_z^i = 0$ and $Z_i = K$ for all $i \in [m]$. Consequently, the regret bound of UCB-ADLO reduces to $\widetilde{O}(\Lambda^i(\lambda_*)K^{1/2}T^{2/3})$ for the $i$-th objective. This improves upon the $\widetilde{O}(K^{2/3}T^{2/3})$ bound in lexicographic MOMAB (Hüyük and Tekin, 2021). Moreover, our regret definition (general regret) is more precise than the priority-based regret used in Hüyük and Tekin (2021).

### 6.2 Analysis

In this section, we provide the proof of Theorem 4. To begin, we first provide the following two lemmas for the UCB framework.

**Lemma 10** *With probability at least* $1-\delta$, *for* $i \in [2]$ *and any* $k \in [\widetilde{K}]$,

$$\left|Y^i(k) - \mathrm{E}[Y^i(k)]\right| \leq 2\sqrt{\ln(8\widetilde{K}/\delta)}.$$

**Proof.** ADLO is a deterministic algorithm, meaning that for a fixed lexicographic bandit problem with given input parameters $\delta, T$ and $\lambda$, the expected accumulated payoffs on executing ADLO is fixed. Therefore, we can take the selection of the trade-off parameters as a stochastic bandit problem, where $[Y^1(k), Y^2(k)]$ represents the payoffs obtained by choosing $\lambda_k$. The proof follows the same reasoning as Lemma 1, but with the candidate arm set $[\widetilde{K}]$. Substituting $m, N_c(r)$ and $n_t(x)$ of Lemma 1 with 2, $\widetilde{K}$ and 1 concludes the proof of Lemma 10. ■

**Lemma 11** *Let* $k' \in [\widetilde{K}]$ *the best trade-off parameter approximating the local trade-off* $\lambda_*$, *such that* $\lambda_{k'}/2 < \lambda_* \leq \lambda_{k'}$. *Then, with probability at least* $1-\delta$, $k' \in \mathcal{K}^1$.

**Proof.** $Y^1(k')$ is obtained by running ADLO with with confidence parameter $1/H$ and trade-off parameter $\lambda_k$. Since $\lambda_* \leq \lambda_{k'}$, according to Corollary 2, we have

$$H\mu^1(x_*) - \mathrm{E}[Y^1(k')] \leq 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}},$$

where $\bar{\alpha}_H = 12\ln(2H)$. Due to the fact[4] that $\mathrm{E}[Y^1(\widetilde{K})] \leq H\mu^1(x_*)$, it follows that

$$\mathrm{E}[Y^1(\widetilde{K})] - \mathrm{E}[Y^1(k')] \leq 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}}.$$

Applying Lemma 10, we have that with probability at least $1-\delta$,

$$Y^1(\widetilde{K}) - Y^1(k') \leq 4\sqrt{\ln(8\widetilde{K}/\delta)} + 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}}.$$

Referring to the Step 7 of UCB-ADLO, we have that with probability at least $1-\delta$, $k' \in \mathcal{K}^1$. The proof of Lemma 11 is finished. ■

[4] This fact does not hold for $i \geq 2$, thus UCB-ADLO can only handle the case $m = 2$.

**Lemma 12** *Let $k' \in [\widetilde{K}]$ the best trade-off parameter to approximate the local trade-off $\lambda_*$, such that $\lambda_{k'}/2 < \lambda_* \leq \lambda_{k'}$. With probability at least $1-\delta$, for $i \in [2]$,*

$$\mathrm{E}[Y^i(k')] - \mathrm{E}[Y^i(k^\dagger)] \leq 8\sqrt{\ln(8\widetilde{K}/\delta)} + 9216\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} H^{\frac{d_z^i+1}{d_z^i+2}}.$$

**Proof.** For the first objective, we decompose $\mathrm{E}[Y^1(k')] - \mathrm{E}[Y^1(k^\dagger)]$ as follows,

$$\mathrm{E}[Y^1(k')] - \mathrm{E}[Y^1(k^\dagger)] \leq H\mu^1(x_*) - \mathrm{E}[Y^1(\widetilde{K})] + \mathrm{E}[Y^1(\widetilde{K})] - \mathrm{E}[Y^1(k^\dagger)]. \tag{33}$$

$Y^1(\widetilde{K})$ is obtained by running ADLO with with confidence parameter $1/H$ and trade-off parameter $\lambda_{\widetilde{K}}$. Since $\lambda_* \leq \lambda_{\widetilde{K}}$, according to Corollary 2, we have

$$H\mu^1(x_*) - \mathrm{E}[Y^1(\widetilde{K})] \leq 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}}. \tag{34}$$

where $\bar{\alpha}_H = 12\ln(2H)$. Next, Lemma 10 shows that with probability at least $1-\delta$,

$$\mathrm{E}[Y^1(\widetilde{K})] - \mathrm{E}[Y^1(k^\dagger)] \leq Y^1(\widetilde{K}) - Y^1(k^\dagger) + 4\sqrt{\ln(8\widetilde{K}/\delta)}. \tag{35}$$

Due to $k^\dagger \in \mathcal{K}^1$, we can further relax Eq. (35) as

$$\mathrm{E}[Y^1(\widetilde{K})] - \mathrm{E}[Y^1(k^\dagger)] \leq 8\sqrt{\ln(8\widetilde{K}/\delta)} + 4608\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}}. \tag{36}$$

Taking Eq. (34) and Eq. (36) into Eq. (33), we have

$$\mathrm{E}[Y^1(k')] - \mathrm{E}[Y^1(k^\dagger)] \leq 8\sqrt{\ln(8\widetilde{K}/\delta)} + 9216\bar{\alpha}_H 2^{d_z^1+2} Z_1^{\frac{1}{d_z^1+2}} H^{\frac{d_z^1+1}{d_z^1+2}}.$$

The proof for the first objective is finished. Moving to the second objective, Lemma 10 tells that

$$\mathrm{E}[Y^2(k')] - \mathrm{E}[Y^2(k^\dagger)] \leq Y^2(k') - Y^2(k^\dagger) + 4\sqrt{\ln(8\widetilde{K}/\delta)}. \tag{37}$$

Meanwhile, Lemma 11 shows that $k' \in \mathcal{K}^1$. Since $k^\dagger = \mathrm{argmax}_{k\in\mathcal{K}^1} Y^2(k^\dagger)$, it follows that $Y^2(k') - Y^2(k^\dagger) \leq 0$. Therefore, Eq. (37) can be relaxed as

$$\mathrm{E}[Y^2(k')] - \mathrm{E}[Y^2(k^\dagger)] \leq 4\sqrt{\ln(8\widetilde{K}/\delta)}.$$

The proof of Lemma 12 is finished. ■

We now proceed to finish the proof of Theorem 4. According to the framework of UCB-ADLO, the total regret decomposes into the individual regret of the base learner ADLO. For objective $i \in [m]$,

$$\mathrm{E}[R^i(T)] = \sum_{\tau=1}^{\lceil T/H \rceil} \sum_{t=(\tau-1)H+1}^{\tau H \wedge T} \mu^i(x_*) - \mathrm{E}[\mu^i(x_t)].$$

In the first $\widetilde{K}$ episodes, UCB-ADLO runs ADLO with $\lambda_k$ for each $k \in [\widetilde{K}]$, and then with $\lambda_{k^\dagger}$ in the remaining episodes. Thus, for objective $i \in [m]$,

$$\begin{aligned} \mathrm{E}[R^i(T)] &\leq \widetilde{K}H + \sum_{\tau=\widetilde{K}+1}^{\lceil T/H \rceil} H\mu^i(x_*) - \mathrm{E}[Y^i(k^\dagger)] \\ &= \widetilde{K}H + \sum_{\tau=\widetilde{K}+1}^{\lceil T/H \rceil} H\mu^i(x_*) - \mathrm{E}[Y^i(k')] + \mathrm{E}[Y^i(k')] - \mathrm{E}[Y^i(k^\dagger)]. \end{aligned} \tag{38}$$

According to Corollary 2 and $\lambda_{k'} \geq \lambda_* > \lambda_{k'}/2$, it holds that for objective $i \in [m]$,

$$H\mu^i(x_*) - \mathrm{E}[Y^i(k')] \leq 9216\Lambda^i(\lambda_*)\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} H^{\frac{d_z^i+1}{d_z^i+2}}. \tag{39}$$

Lemma 12 implies that for objective $i \in [m]$,

$$\mathrm{E}[Y^i(k')] - \mathrm{E}[Y^i(k^\dagger)] \leq 8\sqrt{\ln(8\widetilde{K}/\delta)} + 9216\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} H^{\frac{d_z^i+1}{d_z^i+2}}. \tag{40}$$

Taking Eq. (39) and Eq. (40) into Eq. (38) shows that for objective $i \in [m]$,

$$\mathrm{E}[R^i(T)] \leq \widetilde{K}H + 8\sqrt{\ln(8\widetilde{K}/\delta)} \cdot \lceil T/H \rceil + 18432\Lambda^i(\lambda_*)\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} H^{\frac{d_z^i+1}{d_z^i+2}} \cdot \lceil T/H \rceil.$$

Recalling $H = \lfloor T^{2/3} \rfloor$, simplifies the above equation to

$$\mathrm{E}[R^i(T)] \leq \widetilde{K}T^{\frac{2}{3}} + 16\sqrt{\ln(8\widetilde{K}/\delta)} \cdot T^{\frac{1}{3}} + 36864\Lambda^i(\lambda_*)\bar{\alpha}_H 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} T^{\frac{3d_z^i+4}{3d_z^i+6}}.$$

The proof of Theorem 4 is finished. ■

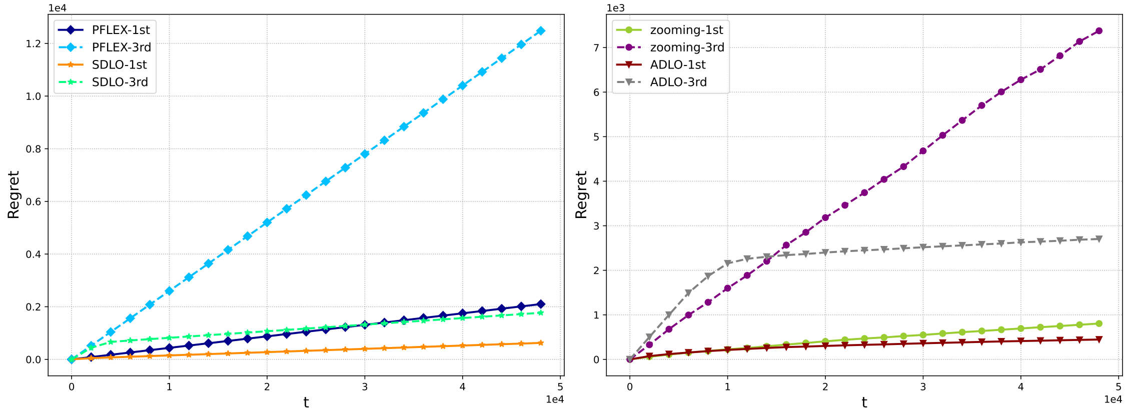
## 7 Experiments

In this section, we perform numerical experiments to validate the efficacy of our algorithms. Our baselines include PF-LEX (Hüyük and Tekin, 2021), a static method tailored for MOMAB with lexicographic ordering, and the zooming algorithm (Kleinberg et al., 2008), an adaptive approach optimized for single-objective Lipschitz bandits.

### 7.1 Parameter-dependent Algorithms

For the parameter-dependent algorithms SDLO and ADLO, we aim to verify two issues through experiments. Firstly, we assess how much improvement our algorithms achieve compared to the baselines (Hüyük and Tekin, 2021; Kleinberg et al., 2008) when the input parameter $\lambda = \lambda_*$. Secondly, we explore the performance of our algorithms when the player has a rough estimation of the trade-off parameter $\lambda_*$.

Following the existing experimental setup (Magureanu et al., 2014), we set the arm space $\mathcal{X} = [0, 1]$ with a Euclidean metric on it. The number of objectives is set as $m = 3$, and the expected payoff functions are given as $\mu^1(x) = 1 - \min_{p \in \{0.1,0.4,0.8\}} 0.5 \times |x - p|$,



(a) Static Methods

(b) Adaptive Methods

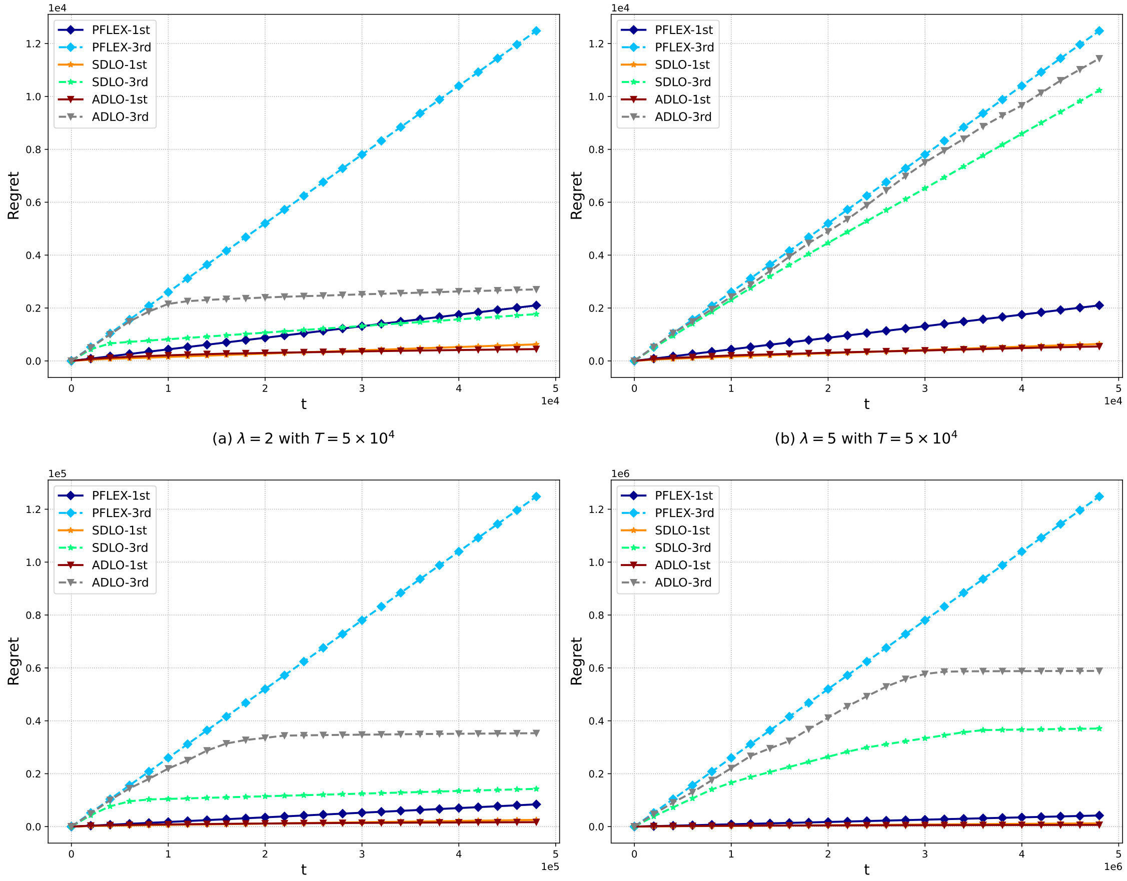
Figure 1: Comparison of our algorithms versus PF-LEX and zooming method with $\lambda = \lambda_*$.

$\mu^2(x) = 1 - \min_{p \in \{0.3, 0.7\}} |x - p|$ and $\mu^3(x) = 1 - |x - 0.4|$, where the local trade-off $\lambda_* = 2$. Note that the optimal arms for the first objective are $\{0.1, 0.4, 0.8\}$, and the optimal arms for both the first and second objectives are $\{0.4, 0.8\}$. Thus, all three objectives have to be considered to determine the lexicographically optimal arm 0.4.

The time horizon $T$ is set to $5 \times 10^4$. According to Corollary 1, the nearly optimal query parameter $r$ for SDLO is 0.025. Consequently, we construct the static arm set for SDLO and PF-LEX as $\mathcal{A} = \{0.025 + 0.05 \times (k - 1) \mid k \in [20]\}$. Meanwhile, PF-LEX requires a parameter $\epsilon$ as input to decide whether to execute exploration or not, we set it to $(|\mathcal{A}|T)^{-\frac{1}{3}} = 0.01$, which is theoretically optimal (Hüyük and Tekin, 2021, Theorem 1). The payoff is obtained by $y_t^i = \mu^i(x_t) + \eta_t$, where $\eta_t$ is drawn from a Gaussian distribution with mean 0 and variance 0.1. To accelerate the convergence rates, the confidence terms of all algorithms are scaled by 0.1, following common practice in bandit learning (Chapelle and Li, 2011; Li et al., 2012; Zhang et al., 2016; Jun et al., 2017a). To reduce randomness, each algorithm is repeated 10 times, and the average regrets of the first and third objectives are reported.

Figure 1(a) illustrates the performance of the static methods, PF-LEX and SDLO. SDLO outperforms PF-LEX in both the first and third objectives. The primary reason for the poor performance of PF-LEX is its theoretically optimal input parameter $\epsilon = 0.01$, which allocates all trials to exploration to ensure that the confidence terms of all arms are smaller than $\epsilon$. Figure 1(b) presents the performance of two adaptive methods: the zooming algorithm and ADLO. ADLO demonstrates comparable regret to the zooming algorithm in the first objective but surpasses it in the third objective. This finding confirms the effectiveness of ADLO in addressing multiobjective problems. The regret curves of ADLO eventually flatten, indicating that it selects the almost optimal arm when $t \geq 10^4$.

In the following, we run our algorithms with different input parameters $\lambda$ to explore how they perform under differing levels of estimation for $\lambda_*$. We adopt the lexicographic MOMAB algorithm PF-LEX as a baseline (Hüyük and Tekin, 2021), which does not require the parameter $\lambda$ as input, instead relying on an exploration parameter $\epsilon$.



(a) $\lambda = 2$ with $T = 5 \times 10^4$

(b) $\lambda = 5$ with $T = 5 \times 10^4$

(c) $\lambda = 5$ with $T = 5 \times 10^5$

(d) $\lambda = 10$ with $T = 5 \times 10^6$

Figure 2: Comparison of our algorithms versus PF-LEX with Different $\lambda$.

Figure 2 presents regret curves for input parameters $\lambda \in \{2, 5, 10\}$. In Figure 2(a), results for $\lambda = 2$ are shown. Here, SDLO and ADLO exhibit comparable performance on the first objective, yet ADLO experiences significantly higher regret on the third objective, contradicting our theoretical guarantees. Specifically, in our experiments, $d_c = 1$ and $d_z^i = 0$ for $i \in [3]$. Corollary 1 and Corollary 2 provide regret bounds of $\widetilde{O}(T^{2/3})$ for SDLO and $\widetilde{O}(T^{1/2})$ for ADLO, respectively. However, ADLO suffers a higher regret on the third objective. This is because the leading constant in Corollary 2 is 2304, which is much larger than the leading constant 32 of Corollary 1. Notably, the lexicographically optimal arm 0.4 is outside the static arm set $\mathcal{A}$, resulting in rising regret curves for SDLO.

In Figure 2(b), we displays the regret curves when the input parameter $\lambda$ of SDLO and ADLO is 5. Comparing Figure 2(a) and Figure 2(b), we observe that for SDLO and ADLO, the regrets for the first objective almost remain unchanged, while the regrets for the third objective increase a lot. This observation is consistent with our theoretical guarantees: for the first objective, $\Lambda^1(\lambda) = 1$, which is independent of $\lambda$, whereas for the third objective, $\Lambda^3(\lambda) = 1 + \lambda + \lambda^2$, which increases with $\lambda$. Meanwhile, for SDLO and ADLO, the regrets of the third objective do not converge. This may because a larger $\lambda$ causes SDLO and ADLO to retain more arms in $\mathcal{A}_s^i$ and $\widetilde{\mathcal{A}}_s^i$ for $i \in \{2,3\}$, requiring SDLO and ADLO to conduct more trials to filter out the promising arms for these two objectives. To verify this analysis, we extended the time horizon to $T = 5 \times 10^5$ in Figure 2(c) and reran SDLO and ADLO with $\lambda = 5$. The regret curves of both SDLO and ADLO nearly flatten, indicating that a larger $\lambda$ increases the number of trials needed to identify the lexicographic optimal arm but does not render the algorithm ineffective.

In Figure 2(b), regret curves are presented for $\lambda = 5$. Comparing with Figure 2(a), we observe consistent regret levels for the first objective in both SDLO and ADLO, while regrets for the third objective notably increase. This aligns with our theoretical expectations: $\Lambda^1(\lambda) = 1$, which is independent of $\lambda$, whereas $\Lambda^3(\lambda) = 1+\lambda+\lambda^2$ grows with $\lambda$. Furthermore, the regrets for the third objective in SDLO and ADLO do not converge, possibly due to a larger $\lambda$ necessitating more trials to identify promising arms for objectives 2 and 3. To validate this, we extended the horizon to $T = 5\times 10^5$ in Figure 2(c) and re-evaluated SDLO and ADLO with $\lambda = 5$, resulting in nearly flat regret curves, indicating that a larger $\lambda$ increases the number of trials needed to identify the lexicographic optimal arm but does not render the algorithm ineffective. Finally, we run SDLO and ADLO with the input parameter $\lambda = 10$ and observe flattened regret curves after $t \geq 3.5 \times 10^6$, reinforcing the conclusion that higher $\lambda$ values do not hinder SDLO and ADLO's effectiveness.

An interesting phenomenon observed is that PF-LEX appears to incur linear regret. This behavior arises because PF-LEX eliminates the optimal arm during the filtering process. Recall that the discretized arm set is defined as $\mathcal{A} = \{0.025+0.05\times(k-1) \mid k \in [20]\}$, where the arm $0.375 \in \mathcal{A}$ is lexicographically optimal. In this setup, the adjacent arm payoff gaps are 0.025 for the first objective and 0.05 for the second. Consider the optimal arm 0.375 and a suboptimal arm 0.325:

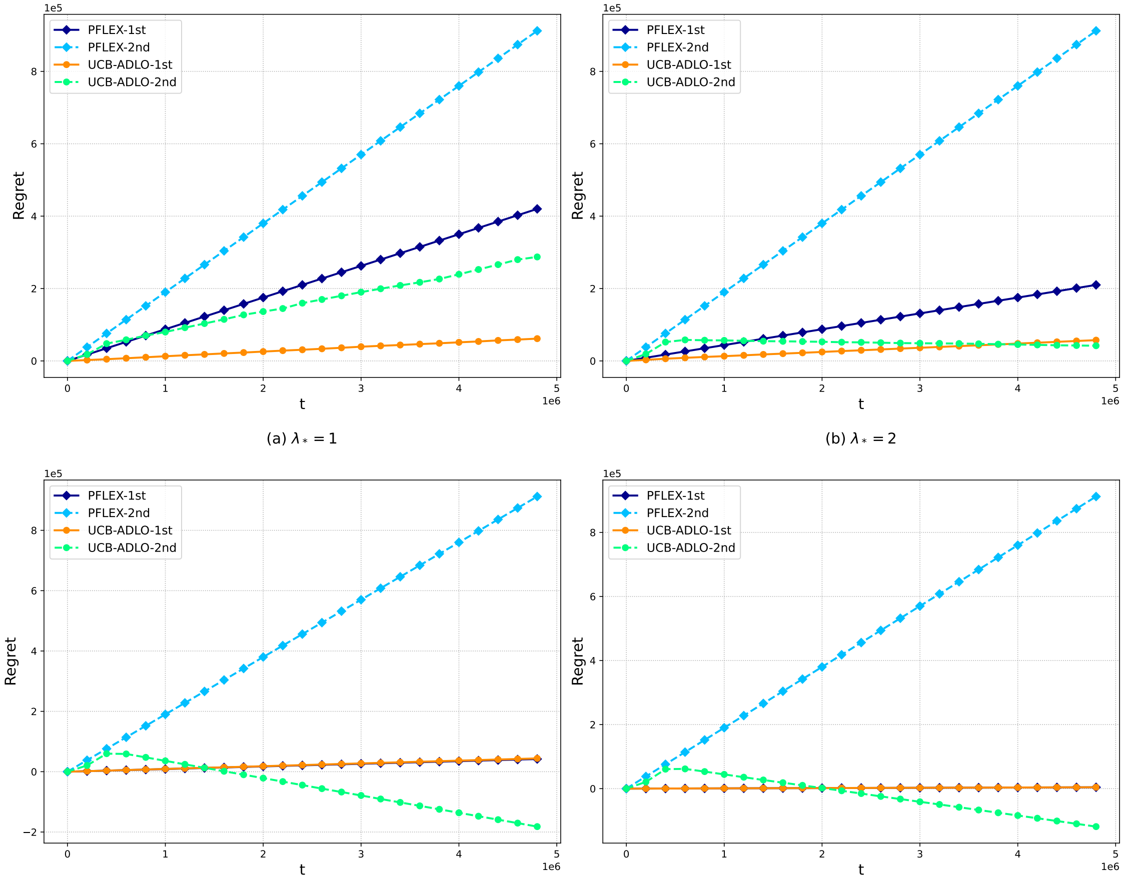
- Their confidence intervals for the first objective overlap, since $\mu^1(0.375)-\mu^1(0.325) = 0.025 < 4\epsilon = 0.04$, allowing arm 0.325 to survive the first-round filtering.
- However, their intervals for the second objective do not overlap, as $\mu^2(0.325) - \mu^2(0.375) = 0.05 > 4\epsilon = 0.04$, which results in the elimination of arm 0.375.

As a consequence, the lexicographically optimal arm is removed from the candidate arms, leading PF-LEX to follow a linear regret curve.

### 7.2 Parameter-free Algorithms

To verify the effectiveness of the parameter-free algorithm UCB-ADLO, we design multiple lexicographic bandit problems, whose local trade-off parameters are different, i.e. $\lambda_* \in \{1, 2, 10, 100\}$. Due to UCB-ADLO can only handle two-objective problems, we set $m = 2$. The expected payoff functions are given as $\mu^1(x) = 1 - \min_{p\in\{0.1,0.4\}} |x-p|/\lambda_*$, $\mu^2(x) = 1 - |x-0.3|$. The other experimental settings are consistent with those described in Section 7.1.

Figure 3 displays the regret curves for PF-LEX and UCB-ADLO. In Figure 3(a), with $\lambda_* = 1$, UCB-ADLO outperforms PF-LEX in both primary and secondary objectives. The inferior performance of PF-LEX can be attributed to exhausting all trials in ensuring that the confidence terms for all arms are smaller than $\epsilon$. In contrast to the regret curves of ADLO shown in Figure 2, those of UCB-ADLO do not flatten. This is due to UCB-ADLO



(a) $\lambda_* = 1$

(b) $\lambda_* = 2$

(c) $\lambda_* = 10$

(d) $\lambda_* = 100$

Figure 3: Comparison of UCB-ADLO versus PF-LEX with Different $\lambda_*$.

restarting the base-learner ADLO every $H = \lfloor T^{2/3} \rfloor$ rounds, necessitating that the regret curves increase every 29,240 rounds.

Moving to Figure 3(b) with $\lambda_* = 2$, two differences from Figure 3(a) emerge. Firstly, the regret of PF-LEX's primary objective decreases, owing to $\mu^1(x) = 1 - \min_{p \in 0.1, 0.4} |x - p| / \lambda_*$. The payoff gap between different arms becomes smaller when divided by $\lambda_* = 2$. Secondly, the regret curve of UCB-ADLO's secondary objective shows a declining trend, attributable to UCB-ADLO selecting $k^\dagger \in \mathcal{K}^1$ such that $\lambda_{k^\dagger} < \lambda_*$, leading $R^2(H) \leq 0$ when running ADLO with the trade-off parameter $\lambda_{k^\dagger}$. The regret curves in Figure 3(c) and (d) resemble those in Figure 3(b), with reduced regrets in PF-LEX's primary objective and accelerated decline in UCB-ADLO's secondary objective. Notably, in Figures 3(b), (c), and (d), the regret for the secondary objective falls significantly below the upper bound constructed in Theorem 4, suggesting potential for further reduction in UCB-ADLO's regret upper bound, a promising area for future exploration.

## 8 Further Discussion

In this section, we discuss the trade-off parameter $\lambda_*$ from four viewpoints: its connection to the global trade-off, its link with the Pareto front, its relationship to the Lipschitz constants of reward functions and its role in regret reduction.

### 8.1 Global Trade-off and $\lambda_*$

Initially, we present an idea known as global trade-off, frequently employed in the field of multiple objective optimization (Miettinen, 1999, Definition 2.8.5). For any $x \in \mathcal{X}$ and $i, j \in [m]$, let $S_{ij}(x) = \{x' \in \mathcal{X} | \mu^j(x') > \mu^j(x), \mu^i(x') \leq \mu^i(x)\}$. The **global trade-off** between $\mu^i(x)$ and $\mu^j(x)$ is defined as

$$\lambda_{ij}(x) = \max_{x' \in S_{ij}(x)} \frac{\mu^i(x) - \mu^i(x')}{\mu^j(x') - \mu^j(x)}.$$

If $S_{ij}(x) = \emptyset$, then $\lambda_{ij}(x) = -\infty$ for $i \neq j$.

For a given arm $x \in \mathcal{X}$, it holds that $\mu^j(x') - \mu^j(x) > 0$ for any $x' \in S_{ij}(x)$, leading to the inequality

$$\mu^i(x) - \mu^i(x') \leq \lambda_{ij}(x) \cdot (\mu^j(x') - \mu^j(x)).$$

Recalling the local trade-off $\lambda_*$ in Eq. (3), we have that

$$\mu^i(x) - \mu^i(x_*) \leq \lambda_* \cdot \max_{j \in [i-1]} \{\mu^j(x_*) - \mu^j(x)\}. \tag{41}$$

Both $\lambda_{ij}(x)$ and $\lambda_*$ characterize the balance among different arms concerning various objectives. Specifically, $\lambda_{ij}(x)$ indicates the balance between arm $x$ and other arms in $S_{ij}(x)$, while $\lambda_*$ represents the balance between arm $x$ and the lexicographically optimal arm $x_*$. Hence, the newly defined parameter $\lambda_*$ is called "local trade-off". In terms of objectives, $\lambda_{ij}(x)$ denotes the balance between objective $i$ and objective $j$, whereas $\lambda_*$ represents the balance between objective $i$ and the top $i-1$ higher-priority objectives. It is important to note that $\lambda_*$ is bounded by $\max_{i>j} \max_{x \in \mathcal{X}} \lambda_{ij}(x)$. Therefore, our parameter-dependent algorithms SDLO and ADLO are applicable to any multiobjective optimization scenario that provides insights into the global trade-off, as long as the input parameter $\lambda$ is set to $\max_{i>j} \max_{x \in \mathcal{X}} \lambda_{ij}(x)$.

### 8.2 Pareto Front and $\lambda_*$

Another noteworthy point is that $\lambda_*$ depends solely on the Pareto front of the multiobjective bandit problem. To clarify this, we define the Pareto order and Pareto front (Li et al., 2021).

Let $u, v \in \mathbb{R}^m$ be two vectors. $u$ is said to **Pareto dominate** $v$ if and only if $\forall i \in [m]$, $u^i \geq v^i$and $\exists i^* \in [m], u^{i^*} > v^{i^*}$. An arm $x \in \mathcal{X}$ is said to be Pareto optimal if and only if its expected payoff is not dominated by any other arm in $\mathcal{X}$. The expected payoffs of all Pareto optimal arms form the **Pareto front**.

If an arm $x$ is not Pareto optimal, there exists a Pareto optimal arm $\tilde{x}$ such that $\exists i \in [m], \mu^i(\tilde{x}) > \mu^i(x)$ and $\forall j \in [m], \mu^j(\tilde{x}) \geq \mu^j(x)$. This leads to the conclusion that

$$\frac{\mu^i(x) - \mu^i(x_*)}{\max_{j \in [i-1]}\{\mu^j(x_*) - \mu^j(x)\}} < \frac{\mu^i(\tilde{x}) - \mu^i(x_*)}{\max_{j \in [i-1]}\{\mu^j(x_*) - \mu^j(\tilde{x})\}} \leq \lambda_*.$$

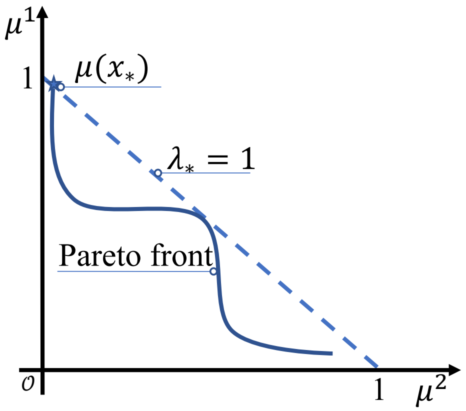


Figure 4: Illustration of $\lambda_*$

Consequently, only the Pareto optimal arm $\tilde{x}$ contribute to determining $\lambda_*$. As presented in Figure 4(a), $\lambda_*$ acts as a parameter describing the relative change of $\mu^1$ and $\mu^2$ on the Pareto front.

### 8.3 Lipschitz Constant and $\lambda_*$

Next, we discuss the relationship between the Lipschitz constant and the local trade-off parameter $\lambda_*$. We clarify this relationship using the two-objective case, where the local trade-off is defined as follows:

$$\begin{aligned}\lambda_* &= \min\left\{\lambda \geq 0 \;\middle|\; \mu^2(x) - \mu^2(x_*) \leq \lambda \cdot [\mu^1(x_*) - \mu^1(x)], \forall x \in \mathcal{X}\right\} \\ &= \min\left\{\lambda \geq 0 \;\middle|\; \frac{\mu^2(x) - \mu^2(x_*)}{\mu^1(x_*) - \mu^1(x)} \leq \lambda, \forall x \in \mathcal{X}\right\} \text{(assuming } \mu^1(x_*) > \mu^1(x), \forall x \neq x_*). \end{aligned} \tag{42}$$

Clearly, $\lambda_*$ is the smallest upper bound on the ratio $\frac{\mu^2(x)-\mu^2(x_*)}{\mu^1(x_*)-\mu^1(x)}$ over all $x \in \mathcal{X}$. To analyze the possible values of $\lambda_*$, it is necessary to derive an upper bound for the numerator $\mu^2(x) - \mu^2(x_*)$ and a lower bound for the denominator $\mu^1(x_*) - \mu^1(x)$.

For the upper bound of $\mu^2(x) - \mu^2(x_*)$, the Lipschitz property of $\mu^2(x)$ provides a sufficient condition, given by:

$$|\mu^2(x) - \mu^2(x_*)| \leq L_2 \cdot \mathcal{D}(x, x_*), \forall x \in \mathcal{X}, L_2 > 0. \tag{43}$$

For the lower bound of $\mu^1(x_*) - \mu^1(x)$, we consider the Lipschitz property of the inverse function of $\mu^1(x)$. Let $\mathcal{X} = [0,1]$, and denote the inverse of $\mu^1(\cdot)$ as $\breve{\mu}^1(\cdot)$. For $y_* = \mu^1(x_*)$ and $y = \mu^1(x)$, the Lipschitz property of $\breve{\mu}^1(\cdot)$ implies

$$|\breve{\mu}^1(y) - \breve{\mu}^1(y_*)| \leq L_1 \cdot |y - y_*|, \forall y \in \mathbb{R}, L_1 > 0.$$

This leads to the inequality:

$$\mathcal{D}(x, x_*) \leq L_1 \cdot |\mu^1(x) - \mu^1(x_*)|, \forall x \in \mathcal{X}, L_1 > 0.$$

where $\mathcal{D}(x, x_*) = |x - x_*|$. Reformulating this gives the following lower bound for $|\mu^1(x) - \mu^1(x_*)|$:

$$\frac{1}{L_1} \cdot \mathcal{D}(x, x_*) \le |\mu^1(x) - \mu^1(x_*)|, \forall x \in \mathcal{X}, L_1 > 0. \tag{44}$$

Combining the Eq. (43) and Eq. (44), the local trade-off in Eq. (42) can be bounded by:

$$\lambda_* \le L_1 \cdot L_2.$$

Therefore, in two-objective case with $\mathcal{X} = [0, 1]$, if the inverse function of $\mu^1(x)$ satisfies the $L_1$-Lipschitz condition and $\mu^2(x)$ satisfies the $L_2$-Lipschitz condition, then the local trade-off parameter satisfies $\lambda_* \le L_1 \cdot L_2$.

Another common source of confusion is whether $\lambda_*$ can be simply characterized as proportional to $1/\min_{x\in\mathcal{X}} \mu^1(x_*) - \mu^1(x)$. In reality, $\lambda_*$ is determined by the maximum ratio

$$\frac{\mu^2(x) - \mu^2(x_*)}{\mu^1(x_*) - \mu^1(x)}$$

across all suboptimal arms $x \in \mathcal{X}$. Consider a simple two-objective, two-arm example with expected payoffs $(1, 0.5)$ and $(0.99, 0.51)$. In this case, $\mu^1(x_*) - \mu^1(x) = 0.01$, so $1/(\mu^1(x_*) - \mu^1(x)) = 100$, while $\lambda_* = 1$, highlighting a substantial discrepancy between the two quantities.

In summary, we can understand $\lambda_*$ from three dimensions. First, $\lambda_*$ is a ratio that describes the trade-offs among different objectives, and its value is smaller than the global trade-off $\max_{i>j} \max_{x\in\mathcal{X}} \lambda_{ij}(x)$. Second, $\lambda_*$ only depends on the Pareto front of the multi-objective bandit problem, reflecting a local property of Pareto optimal arms. Third, in the two-objective case with $\mathcal{X} = [0, 1]$, $\lambda_*$ can be upper bounded by the product of the Lipschitz constant of the inverse function of $\mu^1(x)$ and the Lipschitz constant of $\mu^2(x)$.

### 8.4 Regret Reduction and $\lambda_*$

In this section, we elaborate on two key techniques that reduce the regret bound from $\widetilde{O}(T^{2/3})$ to $\widetilde{O}(\Lambda^i(\lambda) \cdot T^{1/2})$, where $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ for any $i \in [m]$. These techniques are (i) incorporating side information with $\lambda \ge \lambda_*$ and (ii) adopting a multi-stage decision-making strategy. We first show how side information $\lambda$ facilitates the identification of lexicographically optimal arms, thereby tightening the regret bound, and then explain the motivation behind the multi-stage strategy.

#### 8.4.1 How $\lambda$ Reduces Regret

Figure 4 illustrates how $\lambda_*$ characterizes inter-objective trade-offs. It can be shown that given any $\lambda \ge \lambda_*$, the lexicographically optimal arm can be identified as

$$x_* = \underset{x\in\mathcal{X}}{\operatorname{argmax}} (1 + \lambda)^{m-1}\mu^1(x) + (1 + \lambda)^{m-2}\mu^2(x) + \cdots + (1 + \lambda)\mu^{m-1}(x) + \mu^m(x),$$

which is a maximization of the weighted sum of expected payoffs.

Thus, applying the UCB strategy (Auer et al., 2002) to this weighted sum yields sublinear regret. Specifically, for each round $t \in [T]$ and each arm $x \in \mathcal{X}$, define the lower and upper confidence bounds for objective $i \in [m]$ as

$$\ell_t^i(x) = \hat{\mu}_t^i(x) - c_t(x), \quad u_t^i(\cdot) = \hat{\mu}_t^i(x) + c_t(x),$$

where $\hat{\mu}_t^i(x)$ is the empirical mean, and $c_t(x) = \widetilde{O}\left(\sqrt{1/n_t(x)}\right)$ is the confidence term based on the number of times arm $x$ has been pulled up to round $t$.

According to the UCB strategy, the arm selected at round $t$ is

$$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}} (1+\lambda)^{m-1} u_t^1(x) + (1+\lambda)^{m-2} u_t^2(x) + \cdots + (1+\lambda) u_t^{m-1}(x) + u_t^m(x).$$

Under this rule, the cumulative regret is bounded as follows:

$$(1+\lambda)^{m-1} R^1(T) + (1+\lambda)^{m-2} R^2(T) + \cdots + R^m(T) \leq \sum_{t=1}^{T} \sum_{i=1}^{m} (1+\lambda)^{m-i} \cdot 2c_t(x_t).$$

Taking $c_t = \widetilde{O}\left(\sqrt{\frac{1}{n_t(x)}}\right)$ into above equation,we obtain the bound

$$(1+\lambda)^{m-1} R^1(T) + (1+\lambda)^{m-2} R^2(T) + \ldots + R^m(T) \leq \sum_{i=1}^{m} (1+\lambda)^{m-i} \sum_{x \in \mathcal{X}} \sum_{k=1}^{n_T(x)} \widetilde{O}\left(\sqrt{\frac{1}{k}}\right).$$

Using the bound $\sum_{k=1}^{n_T(x)} \widetilde{O}(1/\sqrt{k}) = \widetilde{O}(\sqrt{n_T(x)})$, we get

$$\begin{aligned}(1+\lambda)^{m-1} R^1(T) + (1+\lambda)^{m-2} R^2(T) + \ldots + R^m(T) &\leq \sum_{i=1}^{m} (1+\lambda)^{m-i} \sum_{x \in \mathcal{X}} \widetilde{O}\left(\sqrt{n_T(x)}\right) \\ &\leq \sum_{i=1}^{m} (1+\lambda)^{m-i} \widetilde{O}\left(\sqrt{|\mathcal{X}|T}\right).\end{aligned}$$

This implies that for any $i \in [m]$,

$$R^i(T) \leq \sum_{j=1}^{m} (1+\lambda)^{i-j} \widetilde{O}(\sqrt{|\mathcal{X}|T}). \tag{45}$$

Importantly, this shows that access to the trade-off parameter $\lambda \geq \lambda_*$ allows even a simple UCB-based approach to achieve a regret bound of $\widetilde{O}(\sqrt{T})$. Thus, the knowledge of $\lambda$ directly leads to improved regret performance over PF-LEX.

### 8.4.2 Motivation of the Multi-Stage Strategy

While the weighted-sum UCB approach achieves improved regret, the bound in Eq. (45) scales with the number of objectives $m$. For instance, when $\lambda = 0$ (i.e., no conflict among objectives), the bound becomes $\widetilde{O}(m\sqrt{T})$, which is worse than the $\widetilde{O}(\sqrt{T})$ regret in single-objective settings (Lattimore and Szepesvári, 2020).

To address this limitation, we adopt the framework of Hüyük and Tekin (2021), which handles the prioritized objective by sequentially eliminating suboptimal arms from the most to the least important objective. This procedure prevents the regret of each objective from compounding across all objectives, thereby removing the multiplicative factor $m$ from the regret bound.

However, PF-LEX (Hüyük and Tekin, 2021) employs a pure exploration phase that entirely ignores exploitation, leading to a suboptimal regret bound even equipped with the trade-off parameter $\lambda$. Precisely, PF-LEX (Hüyük and Tekin, 2021) takes a fixed constant $\epsilon > 0$ as input and considers two cases when selecting arms:

(i) If there exists $x_t \in \mathcal{X}$ such that $c_t(x_t) > \epsilon$, PF-LEX selects $x_t$ for exploration.

(ii) Otherwise, it filters the promising arms from the 1-st to the $m$-th objective and selects an arm from the final filtered set.

The case $c_t(x_t) > \epsilon$ corresponds to a pure exploration step, which entirely ignores exploitation and results in suboptimal regret. To overcome this drawback, we replace the fixed exploration parameter $\epsilon$ with an exponentially decreasing sequence $\{\frac{1}{2}, \frac{1}{2^2}, \dots, \frac{1}{2^s}, \frac{1}{2^{s+1}}, \dots\}$, which gradually shift the algorithm from exploration to exploitation. This adaptive exploration–exploitation balance is incorporated into Algorithms 2 and 4, enabling our methods to achieve more efficient regret performance across all objectives and yielding lower overall regret bounds compared to PF-LEX.

## 9 Conclusion and Future Work

In this paper, we study the MOLB model under lexicographic ordering. First, we introduce the local trade-off parameter $\lambda_*$ to depict the trade-off between different objectives. Based on this, we propose two parameter-dependent algorithms, SDLO and ADLO, both requiring a parameter $\lambda > \lambda_*$ as input. SDLO is straightforward but relies on an oracle, and achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda)T^{(1+d_c)/(2+d_c)})$ for the $i$-th objective, where $i \in [m]$, $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$ and $d_c$ is the covering dimension of arm space. ADLO removes the dependence on oracle and achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda)T^{(1+d_z^i)/(2+d_z^i)})$ for the $i$-th objective, where $i \in [m]$ and $d_z^i$ is the zooming dimension of the $i$-th objective. When the arm set is finite, i.e., $\mathcal{X} = K$, both SDLO and ADLO improve the regret bound of PF-LEX (Hüyük and Tekin, 2021) by a factor of $O((KT)^{1/6})$ since $d_c = 0$ and $d_z^i = 0$ in the $K$-armed bandit setting. However, PF-LEX does not require prior knowledge of $\lambda$. Next, we establish a lower bound for lexicographic MOLB problem, indicating that ADLO is optimal in terms of $\lambda_*$ and $T$. Finally, we present UCB-ADLO, a parameter-free algorithm that does require knowledge about the local trade-off $\lambda_*$. For $m = 2$, UCB-ADLO achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda_*)T^{(3d_z^i+4)/(3d_z^i+6)})$ for any objective $i \in [m]$, which is optimal in terms of $\lambda_*$. All regret bounds in this work are derived under the general regret, which independently evaluates the performance of each objective.

Although UCB-ADLO operates without requiring trade-off information $\lambda_*$, its regret bound is suboptimal concerning the time horizon $T$, and it is limited to handling only two objectives ($m = 2$). Therefore, two challenging open problems remain: developing a parameter-free algorithm that achieves a regret bound of $\widetilde{O}(\Lambda^i(\lambda_*)T^{(1+d_z^i)/(2+d_z^i)})$ for the $i$-th objective, and extending the parameter-free algorithm to any number of objectives.

## Acknowledgments and Disclosure of Funding

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 11215622] and by Natural Science Foundation of China [Project No: 62276223].

## Appendix A. Improved Regret Bounds via a Stronger Assumption

The exponential dependence between the trade-off parameter $\lambda \geq \lambda_*$ and objective index $i$ in Theorems 1 and 2 may induce excessive regret for lower-priority objectives. To address this limitation, we introduce a stronger assumption that enables substantially improved regret bounds through modified algorithms.

**Assumption 1** *There exists $\widetilde{\lambda} \geq 0$ such that for all arms $x \in \mathcal{X}$ and objectives $i \geq 2$,*

$$\mu^i(x) - \mu^i(x_*) \leq \widetilde{\lambda} \cdot (\mu^1(x_*) - \mu^1(x)).$$

This assumption quantifies the maximum relative improvement permitted for secondary objectives when degrading the primary objective. Unlike the trade-off parameter $\lambda_*$ defined in Eq. (3), which remains finite in all MOMAB problems, Assumption 1 may fail in scenarios where no finite $\widetilde{\lambda}$ exists. For instance, consider two arms with expected rewards $[2, 2, 2]$ and $[2, 1, 4]$. Here, $\lambda_* = 2$ is well-defined, but no finite $\widetilde{\lambda}$ satisfies the inequality for all objectives, as the third objective's improvement (4 vs. 2) cannot be bounded proportionally to the unchanged first objective. This demonstrates that Assumption 1 imposes a stricter requirement than the trade-off parameter $\lambda_*$ property.

### A.1 Modified SDLO Algorithm and Analysis

**Modified SDLO.** Under Assumption 1, we replace the filtering step (Step 9) in Algorithm 2 with:

$$\mathcal{A}_s^i = \left\{ x \in \mathcal{A}_s^{i-1} \mid \hat{\mu}^i(x) \geq \hat{\mu}^i(\hat{x}_t^i) - (1 + 2\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot (r + 2 \cdot 2^{-s}) \right\}. \tag{46}$$

This modification yields the following refined regret bound.

**Theorem 5** *Suppose that* (2)*,* (5)*, and Assumption 1 hold. If Modified SDLO is run with $r \geq 0$ and $\widetilde{\lambda}$, then with probability at least $1 - \delta$, the regret of SDLO can be bounded as*

$$R^i(T) \leq 2(1 + \widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot \left( rT + 8\sqrt{\alpha_T N_c(r) T} \right), i \in [m]$$

*where $\alpha_T = 4 \ln (m N_c(r) T / \delta)$.*

**Proof Sketch.** The proof follows a structure similar to that of Theorem 1 (Section 4.2.1), relying on two key lemmas, which are adaptations of Lemmas 2 and 3.

**Lemma 13** *Under Assumption 1 and modified filtering* (46)*, with probability at least $1 - \delta$, for any $x \in \mathcal{A}_s$,*

$$\Delta^i(x) \leq 2(1 + \widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot (r + 2 \cdot 2^{-s+1}), i \in [m].$$

**Lemma 14** *Under Assumption 1 and modified filtering* (46)*, with probability at least* $1-\delta$*, the regret for any* $x \in \mathcal{A}$ *can be bounded as*

$$n_T(x)\Delta^i(x) \leq 2(1+\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot \left(rn_T(x) + 8\sqrt{\alpha_T n_T(x)}\right), i \in [m]$$

*where* $\alpha_T = 4\ln(4mN_c(r)T/\delta)$.

**Proof.** The proof mirrors that of Lemma 3. For brevity, we outline the key steps:

1. If $n_T(x) = 1$, Lemma 14 holds trivially.
2. For $n_T(x) \geq 2$, let $s_T(x)$ denote the stage at which $x$ was last chosen. Then:
$$n_T(x) \leq 2^{s_T(x)+1}\sqrt{\alpha_T n_T(x)}. \tag{47}$$
3. Multiplying $n_T(x)$ on both sides of the inequality in Lemma 13 yields
$$n_T(x)\Delta^i(x) \leq 2(1+\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot \left(rn_T(x) + 2 \cdot 2^{-s_T(x)+1}n_T(x)\right). \tag{48}$$
4. Replacing the second term on the right-hand side of Eq. (48) by Eq. (47) concludes the proof of Lemma 14.

■

Equipped with Lemma 14, we can now complete the proof of Theorem 5. For any objective $i \in [m]$, its regret can be written as

$$R^i(T) = T\mu^i(x_*) - \sum_{t=1}^{T}\mu^i(x_t) = \sum_{x\in\mathcal{A}} n_T(x)\Delta^i(x).$$

Applying Lemma 14 to above equation, we obtain

$$R^i(T) \leq \sum_{x\in\mathcal{A}} 2(1+\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot \left(rn_T(x) + 8\sqrt{\alpha_T n_T(x)}\right)$$

Due to the fact $\sum_{x\in\mathcal{A}} n_T(x) = T$ and Cauchy-Schwarz inequality, we obtain that

$$R^i(T) \leq 2(1+\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot \left(rT + 8\sqrt{\alpha_T N_c(r)T}\right).$$

This concludes the proof of Theorem 5. ■

### A.2 Improved Regret Bound for ADLO

**Modified ADLO.** Under Assumption 1, we replace the filtering step (Step 9) in Algorithm 4 with:

$$\widetilde{\mathcal{A}}_s^i = \left\{x \in \widetilde{\mathcal{A}}_s^{i-1} \mid \hat{\mu}^i(x) \geq \hat{\mu}^i(\hat{x}_t^i) - (3+6\widetilde{\lambda}\mathbb{I}(i \geq 2)) \cdot 2^{-s}\right\}, \tag{49}$$

This leads to the following refined bound.

**Theorem 6** *Suppose that* (2), (5) *and Assumption 1 hold. If Modified ADLO is run with* $\widetilde{\lambda}$, *then with probability at least* $1-\delta$, *for any objective* $i \in [m]$, *its regret can be bounded as*

$$R^i(T) \leq \inf_{r_0 \in (0,1)} \left( (1+\widetilde{\lambda}\mathbb{I}(i \geq 2)) r_0 T + 1152(1+\widetilde{\lambda}\mathbb{I}(i \geq 2))\tilde{\alpha}_T \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} N_z^i(2^{-j}) \cdot 2^j \right),$$

*where* $\tilde{\alpha}_T = 4\ln(4mT^2/\delta)$.

**Proof.** The proof follows that of Theorem 2, with $\Lambda^i(\lambda)$ replaced by $1+\widetilde{\lambda}\mathbb{I}(i \geq 2)$. ■

In summary, the exponential dependence $\Lambda^i(\lambda) = 1+\lambda+\cdots+\lambda^{i-1}$ for $i \in [m]$ appears to be intrinsic in the general lexicographic setting. However, under stronger assumptions that impose tighter coupling between objectives (e.g., Assumption 1), this dependence can be significantly reduced.

## Appendix B. Proof of Corollary 1

According to Theorem 1, we have that with probability at least $1-\delta$, for any objective $i \in [m]$, the regret of SDLO can be bounded as

$$R^i(T) \leq 2\Lambda^i(\lambda)\left(rT + 8\sqrt{\alpha_T N_c(r) T}\right) \tag{50}$$

where $\Lambda^i(\lambda) = 1+\lambda+\cdots+\lambda^{i-1}$ and $\alpha_T = 4\ln(mN_c(r)T/\delta)$.

Recalling the definition of covering dimension $d_c$ in Eq. (7), we have that $N_c(r) \leq Cr^{-d_c}$ for some constant $C > 0$. Taking this inequality into Eq. (50), we obtain that for any $i \in [m]$,

$$R^i(T) \leq 2\Lambda^i(\lambda)\left(rT + 8\sqrt{\alpha_T C T / r^{d_c}}\right).$$

To minimize the right hand side of the above inequality, we take $r = T^{-\frac{1}{2+d_c}}$ and obtain that for any $i \in [m]$,

$$R^i(T) \leq 2\Lambda^i(\lambda)\left(T^{\frac{1+d_c}{2+d_c}} + 8(\alpha_T C)^{\frac{1}{2}} T^{\frac{1+d_c}{2+d_c}}\right) \leq 32\Lambda^i(\lambda)(\alpha_T C)^{\frac{1}{2}} T^{\frac{1+d_c}{2+d_c}}.$$

The proof of Corollary 1 is finished. ■

## Appendix C. Proof of Corollary 2

According to Theorem 2, we have that with probability at least $1-\delta$, for any objective $i \in [m]$, the regret of ADLO can be bounded as

$$R^i(T) \leq \inf_{r_0 \in (0,1)} \left( \Lambda^i(\lambda) r_0 T + 1152\Lambda^i(\lambda)\tilde{\alpha}_T \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} N_z^i(2^{-j}) \cdot 2^j \right) \tag{51}$$

where $\Lambda^i(\lambda) = 1+\lambda+\cdots+\lambda^{i-1}$ and $\tilde{\alpha}_T = 8\ln(2mT/\delta)$.

Recalling the definition of zooming dimension $d_z^i$ in Eq. (10), for any $i \in [m]$, there exists a constant $Z_i > 0$ such that $N_z^i(r) \leq Z_i r^{-d_z^i}$ for any $r > 0$. Taking this inequality into Eq. (51), we obtain that for any $i \in [m]$,

$$R^i(T) \leq \inf_{r_0 \in (0,1)} \left( \Lambda^i(\lambda) r_0 T + 1152 \Lambda^i(\lambda) \tilde{\alpha}_T Z_i \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} 2^{j(d_z^i+1)} \right). \tag{52}$$

The last term of above inequality can be further relaxed as

$$\sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} 2^{j(d_z^i+1)} \leq \sum_{j=0}^{\lfloor \log_2 \frac{1}{r_0} \rfloor} 2^{j(d_z^1+1)} \leq \int_{j=0}^{\log_2 \frac{1}{r_0}+1} 2^{j(d_z^i+1)} \mathrm{d}j \leq \frac{(2/r_0)^{d_z^i+1}}{(d_z^i+1)\ln 2}.$$

Taking this into Eq. (52), we obtain that for any $i \in [m]$,

$$R^i(T) \leq \inf_{r_0 \in (0,1)} \left( \Lambda^i(\lambda) r_0 T + 1152 \Lambda^i(\lambda) \tilde{\alpha}_T 2^{d_z^i+2} Z_i r_0^{-(d_z^i+1)} \right).$$

Taking $r_0 = (Z_i/T)^{1/(2+d_z^i)}$, we obtain that for any $i \in [m]$,

$$R^i(T) \leq \Lambda^i(\lambda) Z_i^{\frac{1}{d_z^i+2}} T^{\frac{d_z^i+1}{d_z^i+2}} + 1152 \Lambda^i(\lambda) \tilde{\alpha}_T 2^{d_z^i+2} Z_i^{\frac{1}{d_z^i+2}} T^{\frac{d_z^i+1}{d_z^i+2}}.$$

The proof of Corollary 2 is finished. ■

## Appendix D. Proof of Lemma 2

Let $\bar{x}_{k^*} \in \mathcal{A}$ be the center of ball used to cover the lexicographically optimal arm $x_* \in \mathcal{X}$, such that $x_* \in B(\bar{x}_{k^*}, r)$. We first prove that the sequentially filtering from the first objective to the $m$-th objective in MSDM-SD does not exclude $\bar{x}_{k^*}$ from $\mathcal{A}_s^m$ and that the arms in $\mathcal{A}_s^m$ are promising for all objectives.

**Lemma 15** *If $\bar{x}_{k^*} \in \mathcal{A}_s^0$, then with probability at least $1-\delta$,*

$$\bar{x}_{k^*} \in \mathcal{A}_s^i, \textit{ and}$$
$$\mu^i(x_*) - \mu^i(x) \leq 2\Lambda^i(\lambda) \cdot (r + 2 \cdot 2^{-s}), i \in [m]$$

*for any $x \in \mathcal{A}_s^m$, where $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$.*

**Proof.** We use the induction method with respect to objective index $i \in [m]$ to prove Lemma 8. For $i = 1$, according to Lemma 1 and the Lipschitz property, we obtain

$$\hat{\mu}_t^1(\bar{x}_{k^*}) \geq \mu^1(\bar{x}_{k^*}) - 2^{-s} \geq \mu^1(x_*) - r - 2^{-s}. \tag{53}$$

Since $r_t(x) \leq 2^{-s}$ for any $x \in \mathcal{A}_s^0$, we deduce

$$\mu^1(x_*) \geq \mu_t^1(\hat{x}_t^1) \geq \hat{\mu}_t^1(\hat{x}_t^1) - 2^{-s}. \tag{54}$$

Combining inequalities (53) and (54), we find that

$$\hat{\mu}_t^1(\bar{x}_{k^*}) \geq \hat{\mu}_t^1(\hat{x}_t^1) - r - 2 \cdot 2^{-s},$$

which indicates $\bar{x}_{k^*} \in \mathcal{A}_s^1$ according to the filtering mechanism of MSDM-SD (Step 9).

Next, reusing the fact that $r_t(x) \leq 2^{-s}$ and the Step 9 of MSDM-SD, we get that for any $x \in \mathcal{A}_s^1$,

$$\mu^1(x) \geq \hat{\mu}_t^1(x) - 2^{-s} \geq \hat{\mu}_t^1(\hat{x}_t^1) - r - 3 \cdot 2^{-s}.$$

Due to $\bar{x}_{k^*} \in \mathcal{A}_s^0$ and $\hat{x}_t^1 = \operatorname{argmax}_{x \in \mathcal{A}_s^0} \hat{\mu}_t^1(x)$, the above inequality can be relaxed as

$$\mu^1(x) \geq \hat{\mu}_t^1(\bar{x}_{k^*}) - r - 3 \cdot 2^{-s}.$$

Then, according Lemma 1 and $r_t(x) \leq 2^{-s}$ for any $x \in \mathcal{A}_s^0$, the above inequality can be further relaxed as

$$\mu^1(x) \geq \mu^1(\bar{x}_{k^*}) - r - 4 \cdot 2^{-s}.$$

Since $x_* \in B(\bar{x}_{k^*}, r)$ and $\mu^1(\cdot)$ satisfies Lipschitz property, we obtain that for any $x \in \mathcal{A}_s^1$,

$$\mu^1(x_*) - \mu^1(x) \leq 2r + 4 \cdot 2^{-s}.$$

We have finished the proof for $i = 1$.

Next, we prove that for $i \geq 2$, if $\bar{x}_{k^*} \in \mathcal{A}_{t,s}^{i-1}$ and

$$\mu^j(x_*) - \mu^j(x) \leq 2\Lambda^j(\lambda) \cdot (r + 2 \cdot 2^{-s}), j \in [i-1]$$

for any $x \in \mathcal{A}_{t,s}^{i-1}$, then $\bar{x}_{k^*} \in \mathcal{A}_{t,s}^i$ and

$$\mu^i(x_*) - \mu^i(x) \leq 2\Lambda^i(\lambda) \cdot (r + 2 \cdot 2^{-s}).$$

for any $x \in \mathcal{A}_{t,s}^i$.

According to the definition of local trade-off in Eq. (3), it is evident that

$$\mu^i(x_*) \geq \mu^i(\hat{x}_t^i) - \lambda \cdot \max_{j \in [i-1]} \{\mu^j(x_*) - \mu^j(\hat{x}_t^i)\}. \tag{55}$$

Then, based on Lemma 1 and Lipschitz property, we obtain

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \mu^i(\bar{x}_{k^*}) - 2^{-s} \geq \mu^i(x_*) - r - 2^{-s}. \tag{56}$$

Combining inequalities (55) and (56), we deduce that

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \mu^i(\hat{x}_t^i) - 2\lambda \cdot \Lambda^{i-1}(\lambda)(r + 2 \cdot 2^{-s}) - r - 2^{-s}.$$

Reusing Lemma 1 and $r_t(\hat{x}_t^i) \leq 2^{-s}$, we derive that

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \hat{\mu}_t^i(\hat{x}_t^i) - (2\lambda \cdot \Lambda^{i-1}(\lambda) + 1) \cdot (r + 2 \cdot 2^{-s})$$

which indicates $\bar{x}_{k^*} \in \mathcal{A}_s^i$ according to the filtering mechanism of MSDM-SD (Step 9).

Reusing the fact that $r_t(x) \leq 2^{-s}$ and the Step 9 of MSDM-SD, we get that for any $x \in \mathcal{A}_s^i$,

$$\mu^i(x) \geq \hat{\mu}_t^i(x) - 2^{-s} \geq \hat{\mu}_t^i(\hat{x}_t^i) - 2^{-s} - (2\lambda \cdot \Lambda^{i-1}(\lambda) + 1) \cdot (r + 2 \cdot 2^{-s}).$$

Due to $\bar{x}_{k^*} \in \mathcal{A}_s^{i-1}$ and $\hat{x}_t^i = \operatorname{argmax}_{x \in \mathcal{A}_s^{i-1}} \hat{\mu}_t^i(x)$, the above inequality can be relaxed as

$$\mu^i(x) \geq \hat{\mu}_t^i(\bar{x}_{k^*}) - 2^{-s} - (2\lambda \cdot \Lambda^{i-1}(\lambda) + 1) \cdot (r + 2 \cdot 2^{-s}).$$

Thus, we obtain that for any $x \in \mathcal{A}_s^i$,

$$\begin{aligned}\mu^i(x) &\geq \mu^i(\bar{x}_{k^*}) - 2 \cdot 2^{-s} - (2\lambda \cdot \Lambda^{i-1}(\lambda) + 1) \cdot (r + 2 \cdot 2^{-s}) \\ &\geq \mu^i(x_*) - 2\Lambda^i(\lambda) \cdot (r + 2 \cdot 2^{-s})\end{aligned}$$

where the second inequality is derived from $x_* \in B(\bar{x}_{k^*}, r)$ and $\Lambda^i(\lambda) = 1 + \lambda \cdot \Lambda^{i-1}(\lambda)$. The proof of Lemma 15 is finished. ■

Now, we are ready to prove the Lemma 2. To employ the Lemma 15, we need to prove that $\bar{x}_{k^*} \in \mathcal{A}_s^0$ for $s \geq 1$. We accomplish this through the induction method with respect to the stage index $s$. For $s = 1$, $\bar{x}_{k^*} \in \mathcal{A}_1^0$ obviously since $\mathcal{A}_1^0 = \mathcal{A}$. Next, for $s \geq 2$, we assume that $\bar{x}_{k^*} \in \mathcal{A}_{s-1}^0$. According to Lemma 15, it follows that $\bar{x}_{k^*} \in \mathcal{A}_{s-1}^m$. Given that $\mathcal{A}_s^0 = \mathcal{A}_{s-1}^m$, we conclude that $\bar{x}_{k^*} \in \mathcal{A}_{t,s}^0$.

Since $\bar{x}_{k^*} \in \mathcal{A}_s^0$, Lemma 15 tells that for any $x \in \mathcal{A}_s^m$,

$$\mu^i(x_*) - \mu^i(x) \leq 2\Lambda^i(\lambda) \cdot (r + 2 \cdot 2^{-s}), i \in [m].$$

Thus, Lemma 2 holds for $s \geq 2$ due to $\mathcal{A}_s = \mathcal{A}_{s-1}^m$. For the case when $s = 1$, it is clear that $\mu^i(x_*) - \mu^i(x) \leq 2\Lambda^i(\lambda) \cdot (r + 2)$. The proof of Lemma 2 is finished. ■

## Appendix E. Proof of Lemma 4

The proof of Lemma 4 is similar to the proof of Lemma 1 but with a candidate arm set $\widetilde{\mathcal{A}}_T$. ADLO picks arms at most $T$ arms, we have $|\widetilde{\mathcal{A}}_T| \leq T$. Replacing $N_c(r)$ of Lemma 1 with $T$ concludes the proof of Lemma 1. ■

## Appendix F. Proof of Lemma 5

Let $\tilde{x}_* \in \widetilde{\mathcal{A}}_t$ be the center of the ball used to cover the lexicographically optimal arm $x_* \in \mathcal{X}$, such that $x_* \in B(\tilde{x}_*, r_t(\tilde{x}_*))$. To initiate the proof of Lemma 5, we demonstrate that sequentially filtering from the first objective to the $m$-th objective in MSDM-AD does not exclude $\tilde{x}_*$ from $\widetilde{\mathcal{A}}_{t,s}^m$ and that the arms in $\widetilde{\mathcal{A}}_{t,s}^m$ are promising for all objectives.

**Lemma 16** *If $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^0$, then with probability at least $1 - \delta$,*

$$\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^i, \textit{ and}$$
$$\mu^i(x_*) - \mu^i(x) \leq 6\Lambda^i(\lambda) \cdot 2^{-s}, i \in [m]$$

*for any $x \in \widetilde{\mathcal{A}}_{t,s}^m$, where $\Lambda^i(\lambda) = 1 + \lambda + \cdots + \lambda^{i-1}$.*

**Proof.** To prove this lemma, we employ the induction method with respect to the objective index $i \in [m]$. For $i = 1$, according to Lemma 1 and the Lipschitz property, we obtain

$$\hat{\mu}_t^1(\tilde{x}_*) + 2r_t(\tilde{x}_*) \geq \mu^1(\tilde{x}_*) + r_t(\tilde{x}_*) \geq \mu^1(x_*). \tag{57}$$

Since $r_t(x) \leq 2^{-s}$ for any $x \in \widetilde{\mathcal{A}}_{t,s}^0$, we deduce

$$\mu^1(x_*) \geq \mu_t^1(\hat{x}_t^1) \geq \hat{\mu}_t^1(\hat{x}_t^1) - 2^{-s}. \tag{58}$$

Combining inequalities (57) and (58), we find that

$$\hat{\mu}_t^1(\tilde{x}_*) \geq \hat{\mu}_t^1(\hat{x}_t^1) - 3 \cdot 2^{-s}$$

which indicates $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^1$ according to the filtering mechanism of MSDM-AD (Step 9).

Next, reusing the fact that $r_t(x) \leq 2^{-s}$ and the Step 9 of MSDM-AD, we get that for any $x \in \widetilde{\mathcal{A}}_{t,s}^1$,

$$\mu^1(x) \geq \hat{\mu}_t^1(x) - 2^{-s} \geq \hat{\mu}_t^1(\hat{x}_t^1) - 4 \cdot 2^{-s}.$$

Due to $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^0$ and $\hat{x}_t^1 = \operatorname{argmax}_{x \in \widetilde{\mathcal{A}}_{t,s}^0} \hat{\mu}_t^1(x)$, the above inequality can be relaxed as

$$\mu^1(x) \geq \hat{\mu}_t^1(\tilde{x}_*) - 4 \cdot 2^{-s}.$$

Recalling inequality (57), we obtain that for any $x \in \widetilde{\mathcal{A}}_{t,s}^1$,

$$\mu^1(x_*) - \mu^1(x) \leq 6 \cdot 2^{-s}.$$

The proof for $i = 1$ is completed.

Next, we prove that for $i \geq 2$, if $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^{i-1}$ and

$$\mu^j(x_*) - \mu^j(x) \leq 6\lambda_j \cdot 2^{-s}, j \in [i-1]$$

for any $x \in \widetilde{\mathcal{A}}_{t,s}^{i-1}$, then $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^i$ and

$$\mu^i(x_*) - \mu^i(x) \leq 6\Lambda^i(\lambda) \cdot 2^{-s}$$

for any $x \in \widetilde{\mathcal{A}}_{t,s}^i$.

According to assumption (3), it is evident that

$$\mu^i(x_*) \geq \mu^i(\hat{x}_t^i) - \lambda \cdot \max_{j \in [i-1]} \{\mu^j(x_*) - \mu^j(\hat{x}_t^i)\}. \tag{59}$$

Then, based on Lemma 1 and Lipschitz property, we obtain

$$\hat{\mu}_t^i(\tilde{x}_*) + 2r_t(\tilde{x}_*) \geq \mu^i(\tilde{x}_*) + r(\tilde{x}_*) \geq \mu^i(x_*). \tag{60}$$

Combining inequalities (59) and (60), we deduce that

$$\hat{\mu}_t^i(\tilde{x}_*) + 2r_t(\tilde{x}_*) \geq \mu^i(\hat{x}_t^i) - 6\lambda \cdot \Lambda^{i-1}(\lambda) \cdot 2^{-s}.$$

Due to Lemma 1 and $r_t(x) \le 2^{-s}$ for all $x \in \widetilde{\mathcal{A}}_{t,s}$, we derive that

$$\hat{\mu}_t^i(\tilde{x}_*) \ge \hat{\mu}_t^i(\hat{x}_t^i) - (3 + 6\lambda \cdot \Lambda^{i-1}(\lambda)) \cdot 2^{-s}$$

which indicates $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^i$ according to the filtering mechanism of MSDM-AD (Step 9).

Reusing the fact that $r_t(x) \le 2^{-s}$ and the Step 9 of MSDM-AD, we get that for any $x \in \widetilde{\mathcal{A}}_{t,s}^i$,

$$\mu^i(x) \ge \hat{\mu}_t^i(x) - 2^{-s} \ge \hat{\mu}_t^1(\hat{x}_t^i) - (4 + 6\lambda \cdot \Lambda^{i-1}(\lambda)) \cdot 2^{-s}$$

Due to $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^{i-1}$ and $\hat{x}_t^i = \operatorname{argmax}_{x \in \widetilde{\mathcal{A}}_{t,s}^{i-1}} \hat{\mu}_t^i(x)$, the above inequality can be relaxed as

$$\mu^i(x) \ge \hat{\mu}_t^i(\tilde{x}_*) - (4 + 6\lambda \cdot \Lambda^{i-1}(\lambda)) \cdot 2^{-s}.$$

Recalling inequality (60) and $\Lambda^i(\lambda) = 1 + \lambda \cdot \Lambda^{i-1}(\lambda)$, we obtain that for any $x \in \widetilde{\mathcal{A}}_{t,s}^i$,

$$\mu^i(x) \ge \mu^i(x_*) - 6\Lambda^i(\lambda) \cdot 2^{-s}.$$

The proof of Lemma 16 is finished. ■

Now, we are ready to prove the Lemma 5. To employ the Lemma 16, we need to prove that $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^0$ for $s \ge 1$. We accomplish this through the induction method with respect to the stage index $s$. For $s = 1$, $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,1}^0$ obviously since $\widetilde{\mathcal{A}}_{t,1}^0 = \widetilde{\mathcal{A}}_t$. Next, for the inductive step when $s \ge 2$, we assume that $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s-1}^0$. According to Lemma 16, it follows that $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s-1}^m$. Given that $\widetilde{\mathcal{A}}_{t,s}^0 = \widetilde{\mathcal{A}}_{t,s-1}^m$, we conclude that $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^0$.

Since $\tilde{x}_* \in \widetilde{\mathcal{A}}_{t,s}^0$, Lemma 16 tells that for any $x \in \widetilde{\mathcal{A}}_{t,s}^m$,

$$\mu^i(x_*) - \mu^i(x) \le 6\Lambda^i(\lambda) \cdot 2^{-s}, i \in [m].$$

Thus, Lemma 5 holds for $s \ge 2$ due to $\widetilde{\mathcal{A}}_{t,s} = \widetilde{\mathcal{A}}_{t,s-1}^m$. For the case when $s = 1$, it is clear that $\mu^i(x_*) - \mu^i(x) \le 6\Lambda^i(\lambda)$. The proof of Lemma 5 is finished. ■

## Appendix G. Proof of Lemma 6

We employ the proof by contradiction to establish Lemma 6. Recalling the definition of the $r$-zooming number, $\widetilde{\mathcal{A}}_j^i$ can be covered by no more than $N_z^i(2^{-j})$ balls with radius $2^{-j}/96$. We will now demonstrate that each of these balls contains at most one arm from $\widetilde{\mathcal{A}}_j^i$. Suppose there exists a ball containing two arms $u, v \in \widetilde{\mathcal{A}}_j^i$. On one hand, we have

$$D(u, v) \le \frac{1}{48 \cdot 2^j}.$$

On the other hand, we assume arm $u$ is added into the arm set $\widetilde{\mathcal{A}}_T$ before arm $v$. Let $t$ be the time when $v$ is added into $\widetilde{\mathcal{A}}_T$. Step 4 of ADLO ensures that

$$D(u, v) > r_t(u). \tag{61}$$

If the last time $u$ played before the $t$-th round occurs at the $t'$-th round and $s'$-th stage, we obtain

$$r_t(u) > \frac{1}{2} r_{t'}(u) > 2^{-s'-1}. \tag{62}$$

Lemma 5 states that

$$\Delta^i(u) \leq 12\Lambda^i(\lambda) \cdot 2^{-s'}, i \in [m]. \tag{63}$$

Combining inequalities (62) and (63) reveals that

$$\Delta^i(u) < 24\Lambda^i(\lambda) \cdot r_t(u).$$

Recalling the definition of $\widetilde{\mathcal{A}}_j^i$ in (21), we deduce

$$r_t(u) > \frac{1}{48 \cdot 2^j},$$

which contradicts (61). Therefore, each of the $N_z^i(2^{-j})$ balls contains at most one arm from $\widetilde{\mathcal{A}}_j^i$, implying

$$|\widetilde{\mathcal{A}}_j^i| \leq N_z^i(2^{-j}).$$

The proof is finished. ■

## Appendix H. Proof of Lemma 8

Recalling the definition of $R_c(T)$, we have

$$\begin{aligned} R_c(T) &= \inf_{r_0 \in (0,1)} \left( r_0 T + \ln T \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} N_c(2^{-j}) \cdot 2^j \right) \\ &\leq \inf_{r_0 \in (0,1)} \left( r_0 T + \ln T \cdot N_c(r_0) \sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} 2^j \right) \end{aligned}$$

where the inequality is due to $N_c(r)$ is non-increasing with respect to $r$. The last term of the above inequality can be bounded as

$$\sum_{j \in \mathbb{N}, 2^{-j} \geq r_0} 2^j = \sum_{j=0}^{\lfloor \log_2 \frac{1}{r_0} \rfloor} 2^j \leq 2^{\lfloor \log_2 \frac{1}{r_0} \rfloor + 1} \leq \frac{2}{r_0}.$$

Combining the above inequalities, we have

$$R_c(T) \leq \inf_{r_0 \in (0,1)} \left( r_0 T + 2 \ln T \cdot \frac{N_c(r_0)}{r_0} \right). \tag{64}$$

Define function $f(r) = \frac{N_c(r)}{r^2}$. Since $f(1) = 1$, $\lim_{r \to 0} f(r) = +\infty$, and $f(r)$ is decreasing on $(0, 1)$, there must exist some $\hat{r} \in (0, 1)$ satisfying

$$f(\hat{r}) \leq T \leq f(\hat{r}/2).$$

Taking $\hat{r}$ into the right hand side of Eq. (64) and considering $f(\hat{r}) \leq T$, we have

$$R_c(T) \leq \hat{r} T (1 + 2 \ln T).$$

Due to $\tilde{r} = \frac{R}{6T\ln T}$ and $R \leq R_c(T)$, we have

$$\tilde{r} \leq \frac{\hat{r}T(1+2\ln T)}{6T\ln T} \leq \frac{\hat{r}}{2} \leq \frac{1}{2}.$$

Next, considering $T \leq f(\hat{r}/2)$, we have

$$T\tilde{r}^2 \leq T(\hat{r}/2)^2 \leq f(\hat{r}/2)\cdot(\hat{r}/2)^2 \leq N_c(\tilde{r}).$$

Due to the fact that for any $r > 0$, $N_c(r) \leq N_p(r)$ (Kleinberg et al., 2008) and $N_p(r) \geq 2$, the proof is finished. ■

## Appendix I. Proof of Lemma 9

To bound the KL divergence from $Q_T^0$ to $Q_T^{\tilde{n}}$, we first analyze the KL divergence between the payoff distributions $\boldsymbol{p}_0 = p_0^1 \times \ldots \times p_0^m$ and $\boldsymbol{p}_{\tilde{n}} = p_{\tilde{n}}^1 \times \ldots \times p_{\tilde{n}}^m$. Recalling the distributions defined in Eq. (26) and Eq. (29), the KL divergence can be calculated as

$$\begin{aligned}
KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) =& \prod_{i=1}^m \frac{\mu_0^i(x)}{\Lambda^i(\lambda)}\tilde{r}\ln\left(\prod_{i=1}^m \frac{\mu_0^i(x)}{\mu_{\tilde{n}}^i(x)}\right) \\
&+ \sum_{j=1}^m \left(1 - \frac{\mu_0^j(x)}{\Lambda^j(\lambda)}\tilde{r}\right) \prod_{i=1, i\neq j}^m \frac{\mu_0^i(x)}{\Lambda^i(\lambda)}\tilde{r}\ln\left(\left(1 - \frac{\mu_0^j(x)}{\Lambda^j(\lambda)}\tilde{r}\right)\prod_{i=1,i\neq j}^m \frac{\mu_0^i(x)}{\mu_{\tilde{n}}^i(x)}\right) \\
&+ \cdots + \\
&+ \prod_{i=1}^m \left(1 - \frac{\mu_0^i(x)}{\Lambda^i(\lambda)}\tilde{r}\right)\ln\left(\prod_{i=1}^m \frac{1-\mu_0^i(x)\tilde{r}/\Lambda^i(\lambda)}{1-\mu_{\tilde{n}}^i(x)\tilde{r}/\Lambda^i(\lambda)}\right),
\end{aligned}$$

which considers all possible values of the payoffs. Since $\frac{\mu_0^1(x)}{\Lambda^1(\lambda)} = \frac{\mu_0^2(x)}{\Lambda^2(\lambda)} = \cdots = \frac{\mu_0^m(x)}{\Lambda^m(\lambda)}$ and $\frac{\mu_{\tilde{n}}^1(x)}{\Lambda^1(\lambda)} = \frac{\mu_{\tilde{n}}^2(x)}{\Lambda^2(\lambda)} = \cdots = \frac{\mu_{\tilde{n}}^m(x)}{\Lambda^m(\lambda)}$, the above KL-divergence can be simplified as

$$\begin{aligned}
KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) =& \left(\frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^m \ln\left(\left(\frac{\mu_0^1(x)}{\mu_{\tilde{n}}^1(x)}\right)^m\right) \\
&+ C_m^1\left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)\left(\frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^{m-1} \ln\left(\left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)\left(\frac{\mu_0^1(x)}{\mu_{\tilde{n}}^1(x)}\right)^{m-1}\right) \\
&+ C_m^2\left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^2\left(\frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^{m-2} \ln\left(\left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^2\left(\frac{\mu_0^1(x)}{\mu_{\tilde{n}}^1(x)}\right)^{m-2}\right) \\
&+ \cdots + \\
&+ \left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^m \ln\left(\left(\frac{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda)}{1-\mu_{\tilde{n}}^1(x)\tilde{r}/\Lambda^1(\lambda)}\right)^m\right).
\end{aligned}$$

By applying a further relaxation, we obtain the upper bound

$$\begin{aligned} KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) \le & 2^m \left(\frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^m \ln\left(\left(\frac{\mu_0^1(x)}{\mu_{\tilde{n}}^1(x)}\right)^m\right) \\ & + 2^m \left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)^m \ln\left(\left(\frac{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda)}{1-\mu_{\tilde{n}}^1(x)\tilde{r}/\Lambda^1(\lambda)}\right)^m\right) \\ \le & 2^m \cdot m \cdot \left[\frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\ln\left(\frac{\mu_0^1(x)}{\mu_{\tilde{n}}^1(x)}\right) + \left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)\ln\left(\frac{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda)}{1-\mu_{\tilde{n}}^1(x)\tilde{r}/\Lambda^1(\lambda)}\right)\right]. \end{aligned}$$

Recalling the definition of $\mu_{\tilde{n}}^i(x)$ and $\mu_0^i(x)$ in Eq. (25) and Eq. (28), we can easily obtain

$$\begin{aligned} & \mu_0^i(x) = \mu_{\tilde{n}}^i(x), \forall x \in \mathcal{X} - S_{\tilde{n}}; \\ & \mu_0^i(x) \le \mu_{\tilde{n}}^i(x) \le \mu_0^i(x) + \frac{1}{8m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda)}{\Lambda^m(\lambda)} \cdot \tilde{r}, \forall x \in S_{\tilde{n}}. \end{aligned}$$

Therefore, for any $x \in \mathcal{X} - S_{\tilde{n}}$, we have $KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) = 0$. For any $x \in S_{\tilde{n}}$, we can relax $KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}})$ as

$$KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) \le 2^m \cdot m \cdot \left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)\ln\left(\frac{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda)}{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda) - \tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda))}\right).$$

Due to $\ln(x) \le x - 1$ for any $x > 0$, we can relax the above inequality as

$$\begin{aligned} KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) \le & 2^m \cdot m \cdot \left(1 - \frac{\mu_0^1(x)}{\Lambda^1(\lambda)}\tilde{r}\right)\left(\frac{\tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda))}{1-\mu_0^1(x)\tilde{r}/\Lambda^1(\lambda) - \tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda))}\right) \\ = & \frac{\tilde{r}^2}{8\Lambda^m(\lambda)} + \frac{(\tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda)))^2}{1-\mu_0^i(x)\tilde{r}/\Lambda^i(\lambda) - \tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda))}. \end{aligned}$$

Since $\mu_0^i(x) \in [\frac{\Lambda^i(\lambda)}{2m \cdot 2^m \cdot \Lambda^m(\lambda)}\tilde{r}, \frac{3\Lambda^i(\lambda)}{4m \cdot 2^m \cdot \Lambda^m(\lambda)}\tilde{r}]$, the above inequality can be further relaxed as

$$KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) \le \frac{\tilde{r}^2}{8} + \frac{(\tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda)))^2}{1 - 7\tilde{r}^2/(8m \cdot 2^m \cdot \Lambda^m(\lambda))}.$$

Lemma 8 tells that $\tilde{r} \le 1/2$, which means $4\tilde{r}^2 \le 1$. Thus, we can relax the above inequality as

$$KL(\boldsymbol{p}_0, \boldsymbol{p}_{\tilde{n}}) \le \frac{\tilde{r}^2}{8} + \frac{(\tilde{r}^2/8)^2}{4\tilde{r}^2 - 7\tilde{r}^2/8} = \frac{13}{100}\tilde{r}^2. \tag{65}$$

After analyzing the payoff distributions, we will take the decision at the $t$-th round into account. Let $KL(\cdot, \cdot|\cdot)$ be the conditional KL divergence. For $t = 1, 2, \ldots, T$, we have

$$\begin{aligned} KL(Q_t^0, Q_t^{\tilde{n}}|h_{t-1}) = & \sum_{h_t \in \Omega^t} Q_t^0(h_t)\ln\left(\frac{Q_t^0(h_t|h_{t-1})}{Q_t^{\tilde{n}}(h_t|h_{t-1})}\right) \\ = & \sum_{h_t \in \Omega^t} Q_t^0(h_t)\ln\left(\frac{Q_t^0(x_t|h_{t-1})}{Q_t^{\tilde{n}}(x_t|h_{t-1})} \cdot \frac{Q_t^0(\boldsymbol{y}_t|x_t, h_{t-1})}{Q_t^{\tilde{n}}(\boldsymbol{y}_t|x_t, h_{t-1})}\right) \\ = & \sum_{h_t \in \Omega^t} Q_t^0(h_t)\ln\left(\frac{Q_t^0(\boldsymbol{y}_t|x_t, h_{t-1})}{Q_t^{\tilde{n}}(\boldsymbol{y}_t|x_t, h_{t-1})}\right) \end{aligned}$$

where the last equality is due to for any algorithm $\mathcal{A}$, the distribution of $x_t$ is fixed once the $h_{t-1}$ is given. Based on the full probability formula, the above equation can be rewritten as

$$\begin{aligned}
KL(Q_t^0, Q_t^{\tilde{n}}|h_{t-1}) &= \sum_{h_{t-1}\in\Omega^{t-1}} \sum_{(x_t,\boldsymbol{y}_t)\in\Omega} Q_t^0(x_t,\boldsymbol{y}_t,h_{t-1}) \ln\left(\frac{Q_t^0(\boldsymbol{y}_t|x_t,h_{t-1})}{Q_t^{\tilde{n}}(\boldsymbol{y}_t|x_t,h_{t-1})}\right) \\
&= \sum_{h_{t-1}\in\Omega^{t-1}} \sum_{(x_t,\boldsymbol{y}_t)\in\Omega} Q_t^0(\boldsymbol{y}_t|x_t,h_{t-1}) Q_t^0(x_t,h_{t-1}) \ln\left(\frac{Q_t^0(\boldsymbol{y}_t|x_t,h_{t-1})}{Q_t^{\tilde{n}}(\boldsymbol{y}_t|x_t,h_{t-1})}\right) \\
&= \sum_{h_{t-1}\in\Omega^{t-1}} \sum_{x_t\in\mathcal{X}} Q_t^0(x_t,h_{t-1}) \sum_{\boldsymbol{y}_t\in\{0,\lambda_1/\tilde{r}\}\times\ldots\times\{0,\lambda_m/\tilde{r}\}} Q_t^0(\boldsymbol{y}_t|x_t) \ln\left(\frac{Q_t^0(\boldsymbol{y}_t|x_t)}{Q_t^{\tilde{n}}(\boldsymbol{y}_t|x_t)}\right) \\
&= \sum_{h_{t-1}\in\Omega^{t-1}} \sum_{x_t\in\mathcal{X}} Q_t^0(x_t,h_{t-1}) KL(\boldsymbol{p}_0,\boldsymbol{p}_{\tilde{n}})
\end{aligned}$$

Taking Eq. (65) into the above equation, we get

$$\begin{aligned}
KL(Q_t^0, Q_t^{\tilde{n}}|h_{t-1}) \leq &\sum_{h_{t-1}\in\Omega^{t-1}} \sum_{x_t\in S_{\tilde{n}}} Q_t^0(x_t,h_{t-1}) \frac{13}{100}\tilde{r}^2 \\
=&\frac{13}{100}\tilde{r}^2 Q_t^0(x_t\in S_{\tilde{n}}).
\end{aligned}$$

By the chain rule of KL divergence, we have

$$KL(Q_T^0, Q_T^{\tilde{n}}) = \sum_{t=1}^{T} KL(Q_t^0, Q_t^{\tilde{n}}|h_{t-1}) \leq \frac{13}{100}\tilde{r}^2 \sum_{t=1}^{T} Q_t^0(x_t\in S_{\tilde{n}}) = \frac{13}{100}\tilde{r}^2 \mathrm{E}_{Q_T^0}[Z_{\tilde{n}}]. \quad (66)$$

Recalling the Eq. (30) that $\mathrm{E}_{Q_T^0}[Z_{\tilde{n}}] \leq T/N$ and $N = \max\{2, \lfloor T\tilde{r}^2\rfloor$, we have

$$\mathrm{E}_{Q_T^0}[Z_{\tilde{n}}] \leq \frac{3}{2}\tilde{r}^{-2}.$$

Taking the above equation into Eq. (66) finishes the proof. ■

## Appendix J. Proof of Lemma 13

Let $\bar{x}_{k^*} \in \mathcal{A}$ be the center of ball used to cover the lexicographically optimal arm $x_* \in \mathcal{X}$, such that $x_* \in B(\bar{x}_{k^*}, r)$. Following the proof structure of Lemma 2, we first establish a counterpart of Lemma 15:

**Lemma 17** *Under Assumption 1 and modified filtering* (46), *if* $\bar{x}_{k^*} \in \mathcal{A}_s^0$, *then with probability at least* $1-\delta$,

$$\begin{gathered}
\bar{x}_{k^*} \in \mathcal{A}_s^i, \textit{ and} \\
\mu^i(x_*) - \mu^i(x) \leq 2(1+\widetilde{\lambda}\mathbb{I}(i\geq 2)) \cdot (r + 2\cdot 2^{-s}), i\in[m]
\end{gathered}$$

*for any* $x \in \mathcal{A}_s^m$, *where* $\Lambda^i(\lambda) = 1+\lambda+\cdots+\lambda^{i-1}$.

**Proof.** For $i = 1$, according to Lemma 1 and the Lipschitz property, we obtain

$$\hat{\mu}_t^1(\bar{x}_{k^*}) \geq \mu^1(\bar{x}_{k^*}) - 2^{-s} \geq \mu^1(x_*) - r - 2^{-s}. \tag{67}$$

Since $r_t(x) \leq 2^{-s}$ for any $x \in \mathcal{A}_s^0$, we deduce

$$\mu^1(x_*) \geq \mu_t^1(\hat{x}_t^1) \geq \hat{\mu}_t^1(\hat{x}_t^1) - 2^{-s}. \tag{68}$$

Combining inequalities (67) and (68), we find that

$$\hat{\mu}_t^1(\bar{x}_{k^*}) \geq \hat{\mu}_t^1(\hat{x}_t^1) - r - 2 \cdot 2^{-s},$$

which indicates $\bar{x}_{k^*} \in \mathcal{A}_s^1$ according to the filtering mechanism in Eq. (46).

Next, reusing the fact that $r_t(x) \leq 2^{-s}$ and Eq. (46), we get that for any $x \in \mathcal{A}_s^1$,

$$\mu^1(x) \geq \hat{\mu}_t^1(x) - 2^{-s} \geq \hat{\mu}_t^1(\hat{x}_t^1) - r - 3 \cdot 2^{-s}.$$

Due to $\bar{x}_{k^*} \in \mathcal{A}_s^0$ and $\hat{x}_t^1 = \operatorname{argmax}_{x \in \mathcal{A}_s^0} \hat{\mu}_t^1(x)$, the above inequality can be relaxed as

$$\mu^1(x) \geq \hat{\mu}_t^1(\bar{x}_{k^*}) - r - 3 \cdot 2^{-s}.$$

Then, according Lemma 1 and $r_t(x) \leq 2^{-s}$ for any $x \in \mathcal{A}_s^0$, the above inequality can be further relaxed as

$$\mu^1(x) \geq \mu^1(\bar{x}_{k^*}) - r - 4 \cdot 2^{-s}.$$

Since $x_* \in B(\bar{x}_{k^*}, r)$ and $\mu^1(\cdot)$ satisfies Lipschitz property, we obtain that for any $x \in \mathcal{A}_s^1$,

$$\mu^1(x_*) - \mu^1(x) \leq 2r + 4 \cdot 2^{-s}.$$

We have finished the proof for $i = 1$.

For $i \geq 2$, Assumption 1 tells that

$$\mu^i(x_*) \geq \mu^i(\hat{x}_t^i) - \widetilde{\lambda} \cdot (\mu^1(x_*) - \mu^1(\hat{x}_t^i)). \tag{69}$$

Then, based on Lemma 1 and Lipschitz property, we obtain

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \mu^i(\bar{x}_{k^*}) - 2^{-s} \geq \mu^i(x_*) - r - 2^{-s}. \tag{70}$$

Combining inequalities (69) and (70), we deduce that

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \mu^i(\hat{x}_t^i) - 2\widetilde{\lambda} \cdot (r + 2 \cdot 2^{-s}) - r - 2^{-s}.$$

Reusing Lemma 1 and $r_t(\hat{x}_t^i) \leq 2^{-s}$, we derive that

$$\hat{\mu}_t^i(\bar{x}_{k^*}) \geq \hat{\mu}_t^i(\hat{x}_t^i) - (2\widetilde{\lambda} + 1) \cdot (r + 2 \cdot 2^{-s})$$

which indicates $\bar{x}_{k^*} \in \mathcal{A}_s^i$ according to the filtering mechanism in Eq. (46).

Reusing the fact that $r_t(x) \leq 2^{-s}$ and Eq. (46), we get that for any $x \in \mathcal{A}_s^i$,

$$\mu^i(x) \geq \hat{\mu}_t^i(x) - 2^{-s} \geq \hat{\mu}_t^i(\hat{x}_t^i) - 2^{-s} - (2\widetilde{\lambda} + 1) \cdot (r + 2 \cdot 2^{-s}).$$

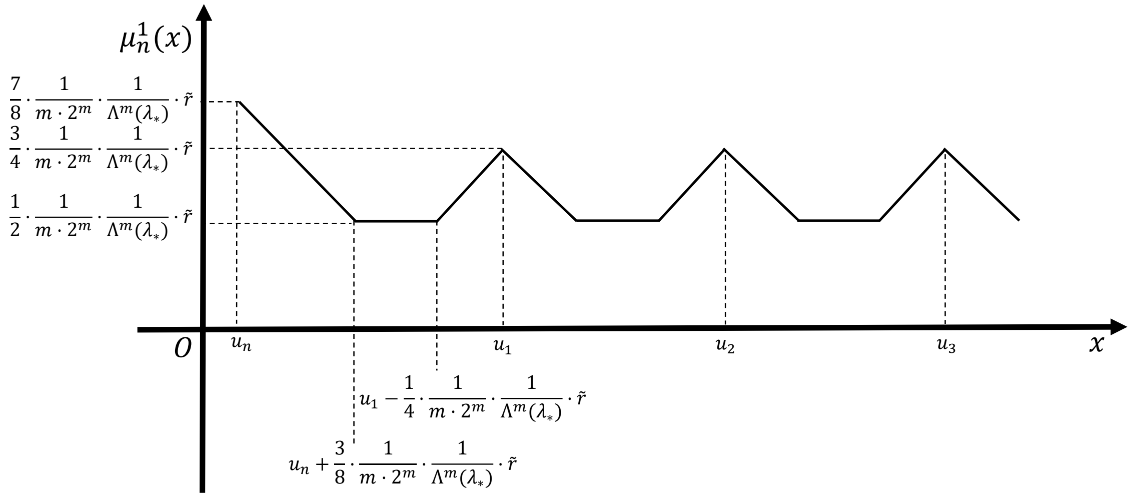


Figure 5: Instances of the Lower Bound.

Due to $\bar{x}_{k^*}\in\mathcal{A}_s^{i-1}$ and $\hat{x}_t^i=\operatorname{argmax}_{x\in\mathcal{A}_s^{i-1}}\hat{\mu}_t^i(x)$, the above inequality can be relaxed as

$$\mu^i(x)\geq\hat{\mu}_t^i(\bar{x}_{k^*})-2^{-s}-(2\widetilde{\lambda}+1)\cdot(r+2\cdot 2^{-s}).$$

Thus, we obtain that for any $x\in\mathcal{A}_s^i$,

$$\begin{aligned}\mu^i(x)&\geq\mu^i(\bar{x}_{k^*})-2\cdot 2^{-s}-(2\widetilde{\lambda}+1)\cdot(r+2\cdot 2^{-s})\\&\geq\mu^i(x_*)-2(\widetilde{\lambda}+1)\cdot(r+2\cdot 2^{-s})\end{aligned}$$

where the second inequality follows from $x_*\in B(\bar{x}_{k^*},r)$. ■

With Lemma 17, the proof of Lemma 13 follows by induction on the stage index $s$, analogous to the proof of Lemma 2. ■

## Appendix K. Proof of the 1-Lipschitz Continuity

In this section, we prove that the constructed instance in the lower bound satisfies the 1-Lipschitz continuity. Recall that for the instance defined in Eq. (25), the expected payoff function for the objective $i\in[m]$ is defined as follows:

$$\mu_n^i(x)=\begin{cases}\frac{7}{8}\cdot\frac{1}{m\cdot 2^m}\cdot\frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)}\cdot\tilde{r}, & x=u_n\\ \frac{3}{4}\cdot\frac{1}{m\cdot 2^m}\cdot\frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)}\cdot\tilde{r}, & x=u_j, j\in[N]\text{ and } j\neq n.\\ \max\left\{\frac{1}{2}\cdot\frac{1}{m\cdot 2^m}\cdot\frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)}\cdot\tilde{r},\max_{u\in U}\mu_n^i(u)-\mathcal{D}(x,u)\right\}, & \text{otherwise}\end{cases}$$

where $U=\{u_1,\ldots,u_N\}\subseteq\mathcal{X}$ denotes the set of designated arms, and $\mathcal{D}(x,y)$ denotes the metric on $\mathcal{X}$.

To aid intuition, Figure 5 illustrates $\mu_n^1(x)$ for the case where $\mathcal{X} = [0,1]$ and $\mathcal{D}(x,y) = |x-y|$. Below, we provide a rigorous proof that $\mu_n^i(x)$ satisfies Lipschitz continuity with Lipschitz constant at most 1. We consider the following cases:

- Case 1: $x, y \in U = \{u_1, \dots, u_N\}$. Let $x = u_j$ and $y = u_k$. The maximum difference in expected payoffs is

  $$|\mu_n^i(x) - \mu_n^i(y)| \leq \frac{1}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}.$$

  Since $\mathcal{D}(x,y) \geq \tilde{r}$ by construction, it follows that

  $$|\mu_n^i(x) - \mu_n^i(y)| \leq \frac{1}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \mathcal{D}(x,y) \leq \mathcal{D}(x,y).$$

- Case 2: $x = u_n$, $\mathcal{D}(u_n, y) \leq \frac{3}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$. In this case, the function is defined as

  $$\mu_n^i(y) = \mu_n^i(x) - \mathcal{D}(x,y),$$

  hence,

  $$|\mu_n^i(x) - \mu_n^i(y)| \leq \mathcal{D}(x,y).$$

- Case 3: $x = u_n$, $\mathcal{D}(u_n, y) > \frac{3}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$, $\forall j \neq n, \mathcal{D}(u_j, y) \geq \frac{1}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$. In this case,

  $$\mu_n^i(y) = \frac{1}{2} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r},$$

  so

  $$|\mu_n^i(x) - \mu_n^i(y)| = \frac{3}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r} < \mathcal{D}(x,y).$$

- Case 4: $x = u_n$, $\mathcal{D}(u_n, y) > \frac{3}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$, $\exists j \in [N], \mathcal{D}(u_j, y) < \frac{1}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r}$. In this case,

  $$\mu_n^i(y) = \frac{3}{4} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r} - \mathcal{D}(u_j, y).$$

  so

  $$|\mu_n^i(x) - \mu_n^i(y)| = \frac{1}{8} \cdot \frac{1}{m \cdot 2^m} \cdot \frac{\Lambda^i(\lambda_*)}{\Lambda^m(\lambda_*)} \cdot \tilde{r} + \mathcal{D}(u_j, y) < \mathcal{D}(x,y).$$

For the case where $x \neq u_n$, the geometric relationship shown in Figure 5 provides the necessary framework to extend the preceding argument, thereby verifying that Eq. (25) satisfies Lipschitz continuity.

## References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995.

Aleks and rs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(73):2533–2568, 2014.

A. D. Athanassopoulos and V. V. Podinovski. Dominance and potential optimality in multiple criteria decision analysis with imprecise information. *The Journal of the Operational Research Society*, 48(2):142–150, 1997.

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

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 454–468, 2007.

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.

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.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems 21*, pages 201–208, 2008.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(46):1655–1695, 2011a.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the lipschitz constant. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, pages 144–158, 2011b.

Sébastien Bubeck, Ofer Dekel, Tomer Koren, and Yuval Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, pages 266–278, 2015.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2011.

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019.

élise crepon, Aurélien Garivier, and Wouter M Koolen. Sequential learning of the Pareto front for multi-objective bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 3583–3591, 2024.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411, 2015.

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.

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

Yasong Feng, Zengfeng Huang, and Tianyu Wang. Lipschitz bandits with batched feedback. In *Advances in Neural Information Processing Systems 35*, pages 19836–19848, 2022.

Yutian Gou, Jinfeng Yi, and Lijun Zhang. Stochastic graphical bandits with heavy-tailed rewards. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 734–744, 2023.

Hadi Hosseini, Sujoy Sikdar, Rohit Vaish, and Lirong Xia. Fair and efficient allocations under lexicographic preferences. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 5472–5480, 2021.

Mark Huber. Halving the bounds for the markov, chebyshev, and chernoff inequalities using smoothing. *The American Mathematical Monthly*, 126(10):915–927, 2019.

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

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.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 99–109, 2017a.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved Strongly Adaptive Online Learning using Coin Betting. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 943–951, 2017b.

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

Ignacy Kaliszewski and Wojtek Michalowski. Searching for psychologically stable solutions of multiple criteria decision problems. *European Journal of Operational Research*, 118(3): 549–562, 1999.

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.

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

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 681–690, 2008.

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.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit Pareto set identification: the fixed budget setting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 2548–2556, 2024.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms.* Cambridge University Press, 2020.

Li Li, Gary G. Yen, Avimanyu Sahoo, Liang Chang, and Tianlong Gu. On the estimation of pareto front and dimensional similarity in many-objective evolutionary algorithm. *Information Sciences*, 563:375–400, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36, 2012.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4154–4163, 2019a.

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, 2019b.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, pages 1739–1776, 2018.

Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 1137–1140, 2018.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of the 27th Conference on Learning Theory*, pages 975–999, 2014.

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

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.

Chara Podimata and Alex Slivkins. Adaptive discretization for adversarial lipschitz bandits. In *Proceedings of the 34st Conference On Learning Theory*, pages 3788–3805, 2021.

Victor V. Podinovski. A dss for multiple criteria decision analysis with imprecisely specified trade-offs. *European Journal of Operational Research*, 113(2):261–270, 1999.

Saba Q. Yahyaa, Madalina M. Drugan, and Bernard Manderick. Knowledge gradient for multi-objective multi-armed bandit algorithms. In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, page 74–83, 2014.

Yuzhen Qin, Yingcong Li, Fabio Pasqualetti, Maryam Fazel, and Samet Oymak. Stochastic contextual bandits with long horizon rewards. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 9525–9533, 2023.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 1952.

Ana B. Ruiz, Francisco Ruiz, Kaisa Miettinen, Laura Delgado-Antequera, and Vesa Ojalehto. NAUTILUS Navigator: free search interactive multiobjective optimization without trading-off. *Journal of Global Optimization*, 74(2):213–231, 2019.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31*, pages 8430–8439, 2018.

Joar Skalse, Lewis Hammond, Charlie Griffin, and Alessandro Abate. Lexicographic multi-objective reinforcement learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 3430–3436, 2022.

Aleksandrs Slivkins. Multi-armed bandits on implicit metric spaces. In *Advances in Neural Information Processing Systems 24*, page 1602–1610, 2011.

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.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

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.

Kristof Van Moffaert, Kevin Van Vaerenbergh, Peter Vrancx, and Ann Nowe. Multi-objective $\mathcal{X}$-armed bandits. In *2014 International Joint Conference on Neural Networks*, pages 2331–2338, 2014.

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.

Yuanyu Wan, Chang Yao, Mingli Song, and Lijun Zhang. Improved regret for bandit convex optimization with delayed feedback. In *Advances in Neural Information Processing Systems 37*, pages 169–196, 2024.

Tianyu Wang, Weicheng Ye, Dawei Geng, and Cynthia Rudin. Towards practical lipschitz bandits. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, page 129–138, 2020.

Nirandika Wanigasekara and Christina Lee Yu. Nonparametric contextual bandits in metric spaces with unknown metric. In *Advances in Neural Information Processing Systems 32*, pages 14657–14667, 2019.

Enrico Weber, Andrea Emilio Rizzoli, Rodolfo Soncini-Sessa, and Andrea Castelletti. Lexicographic optimisation for water resources planning: the case of lake verbano, italy. In *Proceedings of the 1st Biennial Meeting of the International Environmental Modelling and Software Society*, pages 235–240, 2002.

John Myles White. *Bandit Algorithms for Website Optimization*. O'Reilly Media, Inc., 2012.

Kyle Wray, Shlomo Zilberstein, and Abdel-Illah Mouaddib. Multi-objective mdps with conditional lexicographic reward preferences. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, page 3418–3424, 2015.

Kyle Hollins Wray and Shlomo Zilberstein. Multi-objective pomdps with lexicographic reward preferences. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, page 1719–1725, 2015.

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.

Zirui Xu, Xiaofeng Lin, and Vasileios Tzoumas. Bandit submodular maximization for multi-robot coordination in unpredictable and partially observable environments. In *Proceedings of Robotics: Science and Systems*, 2023.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 2936–2942, 2020.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36*, pages 70880–70891, 2023.

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

Bo Xue, Xi Lin, Yuanyu Wan, and Qingfu Zhang. Problem-dependent regret for lexicographic multi-armed bandits with adversarial corruptions. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 6776–6784, 2025a.

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

Sifan Yang, Yuanyu Wan, and Lijun Zhang. Online nonsubmodular optimization with delayed feedback in the bandit setting. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 21992–22000, 2025.

Xiaotian Yu, Han Shao, Michael R. Lyu, and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. *In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, page 937–946, 2018.

Chenxu Zhang, Yibo Wang, Peng Tian, Xiao Cheng, Yuanyu Wan, and Mingli Song. Projection-free bandit convex optimization over strongly convex sets. In *In Proceedings of the 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–129, 2024.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 392–401, 2016.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31*, pages 1323–1333, 2018.

Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021.

Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems 32*, pages 5197–5208, 2019.

Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27574–27590, 2022.