Multiple Trade-offs: An Improved Approach for Lexicographic Linear Bandits

Bo Xue, Xi Lin, Xiaoyuan Zhang, Qingfu Zhang*

Department of Computer Science, City University of Hong Kong, Hong Kong, China
The City University of Hong Kong Shenzhen Research Institute, Shenzhen, China
{boxue4-c, xi.lin, xzhang2523-c}@my.cityu.edu.hk, qingfu.zhang@cityu.edu.hk

Abstract

This paper studies lexicographic online learning within the framework of multiobjective stochastic linear bandits (MOSLB), where the agent aims to simultaneously maximize m objectives in a hierarchical manner. Previous literature has investigated lexicographic online learning in multiobjective multi-armed bandits, a special case of MOSLB. They provided a suboptimal algorithm whose regret bound is $\widetilde{O}(T^{\frac{2}{3}})$ based on a priority-based regret metric. In this paper, we propose an algorithm for lexicographic online learning in the MOSLB model, achieving an almost optimal regret bound of $\widetilde{O}(d\sqrt{T})$ when evaluated by the general regret metric. Here, d is the dimension of arm vectors and T is the time horizon. Our method introduces a new arm filter and a multiple trade-offs approach to effectively balance the exploration and exploitation across different objectives. Experiments confirm the merits of our algorithms and provide compelling evidence to support our analysis.

Introduction

Sequential decision-making under uncertainty arises in numerous real-world applications, including medical trials (Robbins 1952), recommendation systems (Bubeck and Cesa-Bianchi 2012), and autonomous driving (Huang et al. 2019). This has motivated the development of the stochastic multi-armed bandits (MAB). In MAB, an agent repeatedly selects an arm from K arms and receives a single-valued reward sampled from a fixed but unknown distribution specific to the selected arm (Agrawal 1995; Auer 2002; Audibert, Munos, and Szepesvári 2009; Yu et al. 2018; Huang, Dai, and Huang 2022). The goal of the agent is to minimize the regret, which is the cumulative difference between the expected reward of the selected arm and that of the best arm. However, the aforementioned scenarios can be better modeled if multiple objectives are considered. For example, an online advertising system needs to maximize both the click-through rate and the click-conversion rate (Rodriguez, Posse, and Zhang 2012). Therefore, multiobjective multi-armed bandits (MOMAB) model is proposed, which replaces the single-valued reward in MAB with a reward vector (Drugan and Nowe 2013).

MOMAB is a decision-making system that operates over T rounds (Drugan and Nowe 2013). In the t-th round, the agent chooses an arm a_t from the arm set $[K]$[1], and then receives a reward vector $[y_t^1, y_t^2, \ldots, y_t^m] \in \mathbb{R}^m$. Here, m is the number of objectives, and y_t^i represents the reward of the i-th objective, $i \in [m]$. This reward is a random variable with expectation $\mathbb{E}[y_t^i] = \mu^i(a_t)$. Most existing work evaluates the performance of the agent by Pareto regret (Van Moffaert et al. 2014; Turgay, Oner, and Tekin 2018; Lu et al. 2019; Cai et al. 2023; Xu and Klabjan 2023), which regards all objectives as equivalent. Therefore, minimizing the regret of any objective ensures an optimal Pareto regret bound. Specifically, Theorem 4.1 of Xu and Klabjan (2023) states that the Pareto regret is smaller than the regret of any objective $i \in [m]$. Therefore, a nearly optimal Pareto regret bound can be achieved by applying the UCB strategy (Auer 2002) to any of the m objectives. However, the remaining $m-1$ objectives may still suffer the linear regret bounds $O(T)$.

To deal with this inherent drawback, lexicographic order is adopted (Ehrgott 2005), where the priority over m objectives is given by their indices, such that for $i, j \in [m]$, the i-th objective is more important than the j-th objective if and only if $i < j$. When it comes to bandit model, different arms are compared by the lexicographic order on their expected rewards (Hüyük and Tekin 2021). Precisely, given two arms a and a' with expected rewards $[\mu^1(a), \mu^2(a), \ldots, \mu^m(a)]$ and $[\mu^1(a'), \mu^2(a'), \ldots, \mu^m(a')]$, arm a is said to **lexicographically dominate** arm a' if and only if $\mu^1(a) > \mu^1(a')$ or $\exists\ i^* \in \{2, \ldots, m\}$, $\mu^i(a) = \mu^i(a')$ for $1 \leq i \leq i^* - 1$ and $\mu^{i^*}(a) > \mu^{i^*}(a')$. An arm a_* is **lexicographic optimal** if and only if no other arms lexicographically dominate it.

In a recent study, Hüyük and Tekin (2021) investigated the MOMAB problem under lexicographic ordering and proposed a priority-based regret as follows

$$\widehat{R}^i(T) = \sum_{t=1}^{T} \left(\mu^i(a_*) - \mu^i(a_t)\right) \mathbb{I}\left(A^i(a_t)\right), i \in [m]. \quad (1)$$

Here, a_* is the lexicographic optimal arm in $[K]$, $\mathbb{I}(\cdot)$ is the indicator function, and $A^i(a_t)$ is the event that a_t has the same expected rewards as the optimal arm for the previous $i-1$ objectives, i.e., $A^i(a_t) = \{\mu^j(a_*) = \mu^j(a_t), 1 \leq$

*Qingfu Zhang is the corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]We use $[N]$ to denote the set $\{1, 2, \ldots, N\}$.

$j \leq i-1\}$. Hüyük and Tekin (2021) developed an algorithm enjoying a priority-based regret bound of $\widetilde{O}((KT)^{\frac{2}{3}})$.

However, there are three limitations in the work of Hüyük and Tekin (2021). First, the regret bound $\widetilde{O}((KT)^{\frac{2}{3}})$ is suboptimal when applied to the single objective bandit problem since the existing regret bound for single objective MAB is $\widetilde{O}(\sqrt{KT})$ (Bubeck and Cesa-Bianchi 2012). Second, the metric $\widehat{R}^i(T)$ is meaningless for infinite-armed bandits. Specifically, if the number of arms is infinite, applying OFUL[2] to the first objective of a multiobjective bandit problem can find a sequence of arms $\{a_t\}_{t\in[T]}$ that approaches the optimal value of the first objective with a rate of $\widetilde{O}(t^{-\frac{1}{2}})$, but none of $\{a_t\}_{t\in[T]}$ is optimal for the first objective. Therefore, $\mathbb{I}(\mu^1(a_*) = \mu^1(a_t))$ is false for any $t \in [T]$, resulting in the priority-based regret $\widehat{R}^i(T) = 0$ for $i \geq 2$. This is obviously unreasonable because a zero regret indicates OFUL selects the optimal arms for $i \geq 2$ across T rounds, yet OFUL does not optimize the objective $i \geq 2$. Third, MOMAB neglects the contextual information in real-world applications, such as user preferences and news features in news recommendation systems, which could be used to guide the decision-making process (Li et al. 2010).

To remove these limitations, we focus on stochastic linear bandits (SLB), which encodes arms as vectors to incorporate contextual information. Although SLB has been extensively studied in the single objective field (Auer 2002; Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011; Xue et al. 2020; Alieva, Cutkosky, and Das 2021; Zhu and Mineiro 2022; He et al. 2022; Yang et al. 2022), limited research is conducted on the multiobjective setting. In multiobjective stochastic linear bandits (MOSLB), an agent selects an arm $\boldsymbol{x}_t$ from a given arm set $\mathcal{D}_t \subset \mathbb{R}^d$ in the t-th round and then receives a stochastic reward vector $[y_t^1, y_t^2, \ldots, y_t^m] \in \mathbb{R}^m$. The expected reward of each element is linear with the arm vector, i.e.,

$$\mathbb{E}[y_t^i|\boldsymbol{x}_t, \mathcal{F}_{t-1}] = \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t \rangle, i \in [m]. \tag{2}$$

Here, $\boldsymbol{\theta}_*^i$ is the inherent vector for the i-th objective, and $\mathcal{F}_{t-1}$ is a σ-filtration of events up to round t, consisting of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{t-1}\} \cup \{y_1^1, y_2^1, \ldots, y_{t-1}^1\} \cup \ldots \cup \{y_1^m, y_2^m, \ldots, y_{t-1}^m\}$. The stochastic rewards are R-sub-Gaussian for some $R > 0$. In other words, for any $\beta \in \mathbb{R}$,

$$\mathbb{E}[e^{\beta y_t^i}|\boldsymbol{x}_t, \mathcal{F}_{t-1}] \leq \exp\left(\frac{\beta^2 R^2}{2}\right), i \in [m]. \tag{3}$$

We extend the regret of single objective bandits (Lattimore and Szepesvári 2020) to the multiobjective setting, i.e.,

$$R^i(T) = \sum_{t=1}^{T} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t \rangle, i \in [m] \tag{4}$$

where $\boldsymbol{x}_t^*$ is the lexicographic optimal arm in $\mathcal{D}_t$. The existing single objective SLB algorithms exhibit a regret bound of $\widetilde{O}(d\sqrt{T})$ (Dani, Hayes, and Kakade 2008; Abbasi-yadkori, Pál, and Szepesvári 2011). Thus, a compelling challenge for lexicographic MOSLB is to achieve the regret bound of $\widetilde{O}(d\sqrt{T})$ for all m objectives.

To the best of our knowledge, this paper is the first attempt to investigate MOSLB under lexicographic ordering. We extend the metric of lexicographic bandit algorithm from the priority-based regret (1) to the more accurate general regret (4). Subsequently, we develop an algorithm that attains a general regret bound of $\widetilde{O}(d\sqrt{T})$ for all m objectives. This bound is almost optimal in terms of d and T, as the lower bound for the single objective SLB problem is $\Omega(d\sqrt{T})$ (Dani, Hayes, and Kakade 2008). Our algorithm improves upon the study of Hüyük and Tekin (2021), which focused on MOMAB and attained a regret bound of $\widetilde{O}((KT)^{\frac{2}{3}})$. The main innovations of our algorithm include a new arm filter and a multiple trade-offs approach that balance the exploration and exploitation across different objectives. These techniques can be easily adapted to other bandit models, such as finite-armed SLB (Chu et al. 2011), generalized linear bandits (Jun et al. 2017; Xue et al. 2023) and unimodal bandits (Yu and Mannor 2011; Combes and Proutiere 2014).

Related Work

This section provides a literature review on stochastic bandits and multiobjective bandits. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, its ℓ_2-norm is denoted as $\|\boldsymbol{x}\|$, and its induced norm is $\|\boldsymbol{x}\|_V = \sqrt{\boldsymbol{x}^\top V \boldsymbol{x}}$, where $V \in \mathbb{R}^{d\times d}$ is a positive definite matrix.

Stochastic Bandits

The seminal work of Lai and Robbins (1985) not only introduced a stochastic MAB algorithm with a regret bound of $O(K \log T)$ but also established a matching lower bound. Auer (2002) extended the bandit algorithm to the linear model with finite arms and developed the SupLinRel algorithm, which employs a sophisticated device to decouple reward dependence, yielding a regret bound of $\widetilde{O}(\sqrt{dT})$. In the context of infinite-armed stochastic linear bandits, Dani, Hayes, and Kakade (2008) first applied the confidence region technique to deduce an upper confidence bound for the expected rewards, resulting in a regret bound of $\widetilde{O}(d\sqrt{T})$. Meanwhile, Dani, Hayes, and Kakade (2008) provided a matching lower bound $\Omega(d\sqrt{T})$. Later, Abbasi-yadkori, Pál, and Szepesvári (2011) offered a new analysis for the algorithm of Dani, Hayes, and Kakade (2008) and enhanced the regret bound by a logarithmic factor. The most commonly-used strategy for balancing exploration and exploitation in bandit problem is Upper Confidence Bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002; Abbasi-yadkori, Pál, and Szepesvári 2011; Bubeck et al. 2015; Zhang et al. 2016; Hu et al. 2021; Li, Barik, and Honorio 2022; Masoudian, Zimmert, and Seldin 2022; Feng, Huang, and Wang 2022; Jin et al. 2022), which first computes the confidence bound of forthcoming rewards through historical trial data and then selects the arm with the highest upper confidence bound.

To illustrate the UCB technique, we take the classical algorithm OFUL as an example (Abbasi-yadkori, Pál, and

[2]OFUL is a single objective stochastic linear bandit algorithm that achieves a regret bound of $\widetilde{O}(\sqrt{T})$ (Abbasi-yadkori, Pál, and Szepesvári 2011).

Szepesvári 2011). With the trials up to the t-th round, OFUL minimizes the square loss of action-reward pairs $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{t-1}, y_{t-1})\}$ to estimate $\boldsymbol{\theta}_*$, such that,

$$\hat{\boldsymbol{\theta}}_t = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \|X_t \boldsymbol{\theta} - Y_t\|^2 + \|\boldsymbol{\theta}\|^2$$

where $X_t = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_{t-1}^\top] \in \mathbb{R}^{(t-1)\times d}$ is the historical selected arms matrix, and $Y_t = [y_1, y_2, \ldots, y_{t-1}] \in \mathbb{R}^{(t-1)\times 1}$ is historical rewards vector. Based on the estimator $\hat{\boldsymbol{\theta}}_t$, OFUL constructs a confidence region $\mathcal{C}_t$ as follows,

$$\mathcal{C}_t = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|_{V_t} \leq \alpha_t\}$$

where $\alpha_t = O(R\sqrt{d \log(t)})$ and $V_t = I_d + X_t^\top X_t$. Finally, OFUL selects the most promising arm $\boldsymbol{x}_t$ by solving a bilinear optimization problem, i.e.,

$$(\boldsymbol{x}_t, \tilde{\boldsymbol{\theta}}_t) = \underset{\boldsymbol{x} \in \mathcal{D}_t, \boldsymbol{\theta} \in \mathcal{C}_t}{\operatorname{argmax}} \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle. \tag{5}$$

Since the confidence region $\mathcal{C}_t$ is an ellipse, Lagrange method (Boyd and Vandenberghe 2004) shows that the upper confidence bound for the arm $\boldsymbol{x} \in \mathcal{D}_t$ is

$$u_t(\boldsymbol{x}) = \langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{x} \rangle + \alpha_t \|\boldsymbol{x}\|_{V_t^{-1}},$$

where $\langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{x} \rangle$ is an unbiased estimation of $\langle \boldsymbol{\theta}_*, \boldsymbol{x} \rangle$, and $\alpha_t \|\boldsymbol{x}\|_{V_t^{-1}}$ is a confidence term indicating its uncertainty. Thus, we can replace Eq. (5) with $\boldsymbol{x}_t = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{D}_t} u_t(\boldsymbol{x})$.

Multiobjective Bandits

MOMAB was initially studied by Drugan and Nowe (2013), who proposed two UCB-based algorithms that achieve regret bounds of $O(K \log T)$ under the Pareto regret metric and scalarized regret metric, respectively. The Pareto regret measures the cumulative distance between the obtained rewards and the Pareto optimal rewards, and the scalarized regret is the weighted sum of the regrets across all objectives (Drugan and Nowe 2013). To leverage side information, Turgay, Oner, and Tekin (2018) examined the multiobjective contextual bandit model, where the expected reward satisfies the Lipschitz condition. Lu et al. (2019) developed an algorithm for the multiobjective generalized linear bandit model, achieving a Pareto regret bound of $\widetilde{O}(d\sqrt{T})$. Another research direction focuses on designing algorithms from the perspective of best arm identification, whose primary goal is to identify the Pareto optimal arms within a limited budget (Van Moffaert et al. 2014; Auer et al. 2016; Azizi, Kveton, and Ghavamzadeh 2022; Kone, Kaufmann, and Richert 2024). Hüyük and Tekin (2021) studied the lexicographic MOMAB problem and presented an algorithm called PF-LEX, which has a regret bound of $\widetilde{O}((KT)^{\frac{2}{3}})$ under the priority-based regret metric (1). Cheng et al. (2024) investigated the mixed lexicographic-Pareto order for MOSLB problem. Xue et al. (2024) achieved a regret bound of $\widetilde{O}(T^{1-1/(d_z^i+2)})$ for lexicographic Lipschitz bandits, where d_z^i is the zooming dimension of objective $i \in [m]$. However, this regret bound is nearly linear when d_z^i is large.

The intuitive idea to settle the lexicographic bandit problem is to sequentially filter the arms according to the priority among objectives (Ehrgott 2005; Hüyük and Tekin 2021). To further illustrate this, we introduce the PF-LEX algorithm (Hüyük and Tekin 2021). At each round t, PF-LEX first calculates the estimated reward for each arm $a \in [K]$ and objective $i \in [M]$, which is $\hat{\mu}_t^i(a) = \sum_{\tau=1}^{t-1} y_\tau^i \mathbb{I}(a_\tau = a)/N_t(a)$, where a_τ is the arm played at round τ and $N_t(a)$ denotes the number of times arm a has been played up to round t. Based on this, PF-LEX constructs the confidence intervals of the expected rewards as follows

$$\left[\hat{\mu}_t^i(a) - w_t(a), \hat{\mu}_t^i(a) + w_t(a)\right]$$

where $w_t(a) = \beta_t \sqrt{\frac{1+N_t(a)}{N_t^2(a)}}$ and $\beta_t = O(\sqrt{\log(Kmt)})$. Subsequently, PF-LEX either chooses an arm with a wide confidence interval to explore potentially better arms or selects the arm that is most likely optimal in all objectives. Precisely, taking some $\epsilon > 0$ as an input for PF-LEX, if there exists some arm $a_t \in [K]$ such that $w_t(a_t) > \epsilon$, PF-LEX chooses this arm a_t. On the other hand, if $w_t(a) < \epsilon$ for all arms $a \in [K]$, PF-LEX filters the promising arms through the chain relation (Joseph et al. 2016). Starting from $\mathcal{A}_t^0 = [K]$, PF-LEX operates as follows: for $i \in [m]$,

$$\hat{a}_t^i = \underset{a \in \mathcal{A}_t^{i-1}}{\operatorname{argmax}} \hat{\mu}_t^i(a) + w_t(a), \mathcal{A}_t^i = \{a \in \mathcal{A}_t^{i-1} | a C_i \hat{a}_t^i\}.$$

$aC_i\hat{a}_t^i$ denotes that arm a and $\hat{a}_t^i$ are chained, i.e., there exists a sequence of arms $\{a, b_1, \ldots, b_n, \hat{a}_t^i\} \subseteq [K]$, the confidence intervals of adjacent arms are intersected on the i-th objective. Finally, PF-LEX selects the arm $\hat{a}_t^m$.

Algorithms

In this section, we first extend the existing MOMAB algorithm (Hüyük and Tekin 2021) to the MOSLB model as a warm-up, whose regret bound is suboptimal. Then, we introduce a novel arm filter and a multiple trade-offs approach to improve the regret bound to an almost optimal level.

Without loss of generality, we assume that the arm vectors and the inherent vectors are bounded, i.e., for any $t \in [T]$ and $\boldsymbol{x} \in \mathcal{D}_t$, $\|\boldsymbol{x}\| \leq 1$ and for any $i \in [m]$, $\|\boldsymbol{\theta}_*^i\| \leq U$.

Warm-up: STLO

We introduce an algorithm called Single Trade-off under Lexicographic Ordering (STLO), which is an extension of PF-LEX (Hüyük and Tekin 2021), as shown in Algorithm 1.

STLO takes a confidence parameter $\delta \in (0, 1)$ and an exploration parameter $\epsilon > 0$ as input. Before the starting of round t, STLO has collected the historical data from the previous $t-1$ rounds, such as the historical selected arms and the corresponding rewards for all m objectives. These trial data can be used to estimate the unknown vectors $\{\boldsymbol{\theta}_*^i\}_{i \in [m]}$. Specifically, in round t, the estimator for the objective $i \in [m]$ is computed as follows:

$$\hat{\boldsymbol{\theta}}_t^i = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \|X_t \boldsymbol{\theta} - Y_t^i\|^2 + \|\boldsymbol{\theta}\|^2 \tag{6}$$

where $X_t = [\boldsymbol{x}_\tau^\top]_{\tau \in [t-1]} \in \mathbb{R}^{(t-1)\times d}$ is the matrix of historically selected arms, and $Y_t^i = [y_\tau^i]_{\tau \in [t-1]} \in \mathbb{R}^{(t-1)\times 1}$ is the historical rewards of the i-th objective [3].

[3] $\hat{\boldsymbol{\theta}}_1^1 = \hat{\boldsymbol{\theta}}_1^2 = \cdots = \hat{\boldsymbol{\theta}}_1^m = \mathbf{0}$, where $\mathbf{0}$ is the zero vector of $\mathbb{R}^d$.

Algorithm 1: Single Trade-off under Lexicographic Ordering (STLO)

Require: exploration parameter ϵ, time horizon T

for $t = 1, 2, \ldots,$ **do**
  Compute the estimators $\hat{\boldsymbol{\theta}}_t^i$ for all $i \in [m]$ by Eq. (6)
  Compute the estimated rewards and the confidence term for all arms in $\mathcal{D}_t$ by Eq. (7)
  if $w_t(\boldsymbol{x}_t) > \epsilon$ for some $\boldsymbol{x}_t \in \mathcal{D}_t$ **then**
    Play the arm $\boldsymbol{x}_t$ and observe $[y_t^1, y_t^2, \ldots, y_t^m]$
  else
    Initialize $\mathcal{D}_t^0 = \mathcal{D}_t$
    for $i = 1, 2, \ldots, m$ **do**
      $\hat{\boldsymbol{x}}_t^i = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{D}_t^{i-1}} \hat{y}_t^i(\boldsymbol{x})$
      $\mathcal{D}_t^i = \left\{\boldsymbol{x} \in \mathcal{D}_t^{i-1} | \boldsymbol{x} \cdot C_\epsilon^i \cdot \hat{\boldsymbol{x}}_t^i\right\}$
    end for
    Play the arm $\boldsymbol{x}_t = \hat{\boldsymbol{x}}_t^m$ and observe $[y_t^1, \ldots, y_t^m]$
  end if
  Update $X_{t+1} = [\boldsymbol{x}_\tau^\top]_{\tau \in [t]}$, $Y_{t+1}^i = [y_\tau^i]_{\tau \in [t]}, i \in [m]$
end for

Using a variant of the self-normalized bound for martingales (Abbasi-yadkori, Pál, and Szepesvári 2011), we obtain the estimated rewards and the confidence term for any arm $\boldsymbol{x} \in \mathcal{D}_t$ as follows:

$$\hat{y}_t^i(\boldsymbol{x}) = \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle, w_t(\boldsymbol{x}) = \gamma_t \|\boldsymbol{x}\|_{V_t^{-1}}, i \in [m] \quad (7)$$

where $\gamma_t = R\sqrt{d \ln(2mt/\delta)} + U$, $V_t = I_d + X_t^\top X_t$ and I_d is the d-dimensional identity matrix.

Equipped with the confidence terms $\{w_t(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{D}_t}$, STLO divides the decision-making process into two cases based on the exploration parameter ϵ. Precisely, **if** there exists some $\boldsymbol{x}_t \in \mathcal{D}_t$ whose confidence term is larger than ϵ, i.e., $w_t(\boldsymbol{x}_t) > \epsilon$, STLO plays this arm. In this case, STLO engages in pure exploration without any exploitation. **In contrast**, if all arms have reliable estimated rewards, i.e., $w_t(\boldsymbol{x}) \leq \epsilon, \forall \boldsymbol{x} \in \mathcal{D}_t$, an intuitive method is to play the arm with the highest upper confidence bound to make a trade-off between exploration and exploitation. **However**, the arm with the highest upper confidence bound may vary for different objectives, preventing the maximization of all objectives by selecting a single arm. Therefore, STLO filters arms from the first objective to the last objective sequentially as the objectives are ranked by its importance.

Before introducing the filtering mechanism of STLO, we provide the ϵ-chain relation, a simple variant of the chain relation (Hüyük and Tekin 2021). Since the confidence terms are smaller than ϵ in the **else** case, for any arm $\boldsymbol{x} \in \mathcal{D}_t$, its ϵ-confidence intervals for the expected rewards of the i-th objective is constructed as

$$[\ell_t^i(\boldsymbol{x}, \epsilon), u_t^i(\boldsymbol{x}, \epsilon)] = [\hat{y}_t^i(\boldsymbol{x}) - \epsilon, \hat{y}_t^i(\boldsymbol{x}) + \epsilon]. \quad (8)$$

For any arms $\boldsymbol{z}_1, \boldsymbol{z}_n \in \mathcal{D}_t$ and objective $i \in [m]$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_n$ are ϵ-chained with each other on the i-th objective if and only if there exists a sequence of arms $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\} \subseteq \mathcal{D}_t$ whose ϵ-confidence intervals of adjacent arms intersect, i.e., $[\ell_t^i(\boldsymbol{z}_j, \epsilon), u_t^i(\boldsymbol{z}_j, \epsilon)] \cap [\ell_t^i(\boldsymbol{z}_{j+1}, \epsilon), u_t^i(\boldsymbol{z}_{j+1}, \epsilon)] \neq \emptyset, \forall j \in [n-1]$. We use $\boldsymbol{z}_1 \cdot C_\epsilon^i \cdot \boldsymbol{z}_n$ to denote $\boldsymbol{z}_1$ and $\boldsymbol{z}_n$ are ϵ-chained with each other on the i-th objective.

If all arms have reliable estimated rewards, i.e., $w_t(\boldsymbol{x}) \leq \epsilon, \forall \boldsymbol{x} \in \mathcal{D}_t$, STLO filters out the promising arms from the first objective to the last objective. Let $\mathcal{D}_t^0 = \mathcal{D}_t$. For the i-th objective, STLO first selects the arm with the highest estimated reward, such that $\hat{\boldsymbol{x}}_t^i = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{D}_t^{i-1}} \hat{y}_t^i(\boldsymbol{x})$. Then, it filters arms who are ϵ-chained with the arm $\hat{\boldsymbol{x}}_t^i$, i.e.,

$$\mathcal{D}_t^i = \left\{\boldsymbol{x} \in \mathcal{D}_t^{i-1} | \boldsymbol{x} \cdot C_\epsilon^i \cdot \hat{\boldsymbol{x}}_t^i\right\}. \quad (9)$$

After repeating these two steps on all m objectives, STLO plays the arm $\hat{\boldsymbol{x}}_t^m$ and observes the reward $[y_t^1, y_t^2, \ldots, y_t^m]$. Finally, STLO updates the contextual information matrix X_{t+1} and the historical rewards vector Y_{t+1}^i of each objective $i \in [m]$ to prepare for the decision of the next round.

STLO is evaluated by the priority-based regret (1). We first covert it into linear bandit version, i.e.,

$$\widehat{R}^i(T) = \sum_{t=1}^T \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t \rangle \mathbb{I}\left(A^i(\boldsymbol{x}_t)\right)$$

where $A^i(\boldsymbol{x}_t)$ is the event that $\boldsymbol{x}_t$ has the same expected rewards as the optimal arm for the previous $i-1$ objectives, i.e., $A^i(\boldsymbol{x}_t) = \{\langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t^* \rangle = \langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t \rangle, 1 \leq j \leq i-1\}$. Base on this, we can give the upper regret bound for STLO.

Theorem 1 *Suppose that* (2) *and* (3) *hold, and the arm sets are finite, i.e.,* $|\mathcal{D}_t| = K, \forall t \in [T]$. *If STLO is run with* $\epsilon > 0$, *then with probability at least* $1 - \delta$, *for any objective* $i \in [m]$, *its regret satisfies*

$$\widehat{R}^i(T) \leq 50dU \ln T \gamma_T^2 \cdot \epsilon^{-2} + 2KT\epsilon$$

where $\gamma_T = R\sqrt{d \ln(2mT/\delta)} + U$.

Remark 1 Theorem 1 states that STLO achieves a regret bound of $\widetilde{O}(d^2 \cdot \epsilon^{-2} + KT\epsilon)$ for all objectives. Let $\epsilon = d^{\frac{4}{3}}(KT)^{-\frac{1}{3}}$, this bound can be minimized to $\widetilde{O}(d^{\frac{4}{3}}(KT)^{\frac{2}{3}})$. STLO allows the arm set $\mathcal{D}_t$ to be varied for $t \in [T]$, which distinguishes itself from PF-LEX (Hüyük and Tekin 2021). Furthermore, ϵ-chain relation in (9) can be realized with a complexity of $\widetilde{O}(|\mathcal{D}_t|)$, while the chain relation in PF-LEX suffers from a complexity of $O(|\mathcal{D}_t|^2)$. More details of implementing ϵ-chain relation are available in Appendix A.

Although STLO is straightforward, it has two limitations. First, the priority-based regret $\widehat{R}^i(T)$ depends on the indicator function $\mathbb{I}(\cdot)$. Therefore, for any $i \geq 2$, if there exists some $j \in [i-1]$ such that $\langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t^* \rangle > \langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t \rangle$, then the instantaneous regret $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t \rangle$ cannot be accumulated to the total regret. Second, STLO is suboptimal for single objective SLB problems because the lower bound for finite-armed SLB is $\Omega(\sqrt{dT})$ (Chu et al. 2011).

Improved Algorithm: MTLO

In this section, we give an improved algorithm called Multiple Trade-offs under Lexicographic Ordering (MTLO), which removes the indicator function of $\widehat{R}^i(T)$ through a new-designed arm filter and achieves the almost optimal regret bound by employing a multiple trade-offs approach.

Algorithm 2: Lexicographic Ordered Arm Filter (LOAF)

Require: candidate arm set $\mathcal{D} \subseteq \mathcal{D}_t$, confidence width W

1: Initialize the arm set $\mathcal{D}^0 = \mathcal{D}$
2: **for** $i = 1, 2, \ldots, m$ **do**
3: $\quad \hat{\boldsymbol{x}}_t^i = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{D}^{i-1}} \hat{y}_t^i(\boldsymbol{x}) + w_t(\boldsymbol{x})$
4: $\quad \mathcal{D}^i = \{\boldsymbol{x} \in \mathcal{D}^{i-1} | \hat{y}_t^i(\boldsymbol{x}) + w_t(\boldsymbol{x}) \geq \hat{y}_t^i(\hat{\boldsymbol{x}}_t^i) + w_t(\hat{\boldsymbol{x}}_t^i) - (2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1}) \cdot W\}$
5: **end for**
6: **Return** the filtered arm set $\mathcal{D}^m$

Lexicographic Ordered Arm Filter To demonstrate the inspiration behind our proposed arm filter, we first clarify why STLO requires the indicator function $\mathbb{I}(\cdot)$ in its metric. The core issue is that, for any $i \geq 2$, while filtering arms from $\mathcal{D}_t^{i-1}$ to $\mathcal{D}_t^i$, the ϵ-chain relation $\cdot C_\epsilon^i \cdot$ may result in the absence of the lexicographic optimal arm $\boldsymbol{x}_t^*$. To clarify this point, we provide a simple example in Figure 1.

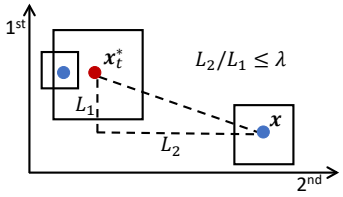

Figure 1: Inspiration

Figure 1 depicts three arms, where the red point represents the lexicographic optimal arm $\boldsymbol{x}_t^*$. The square denotes the confidence intervals for the first and second objectives. Recall STLO (Steps 8-11), $\hat{\boldsymbol{x}}_t^1 = \boldsymbol{x}_t^*$ and $\mathcal{D}_t^1$ contains both $\boldsymbol{x}_t^*$ and $\boldsymbol{x}$ since their confidence intervals for the first objective intersect. However, $\mathcal{D}_t^2$ loses $\boldsymbol{x}_t^*$ because $\hat{\boldsymbol{x}}_t^2 = \boldsymbol{x}$ and $\hat{\boldsymbol{x}}_t^2$ is not chained with $\boldsymbol{x}_t^*$ in the second objective. If $m > 2$, such a case make the subsequent filtering operation for $i > 2$ has no optimal arm for comparison, rendering the theoretical analysis infeasible. Thus, the metric $\widehat{R}^i(T)$ has the indicator function, which is an assumption that the lexicographic optimal arm is not lost during the analysis of the regret bound.

To remove the indicator function, we need to design a filter that dose not lose the optimal arm $\boldsymbol{x}_t^*$. We observe that for a fixed arm, the width of confidence interval is equal among different objectives. Therefore, we can scale the confidence intervals of the objective $i \geq 2$ to ensure the confidence intervals of $\boldsymbol{x}_t^*$ and $\hat{\boldsymbol{x}}_t^i$ are intersected. Inspired by this idea, a trade-off parameter $\lambda \geq 0$ is introduced, such that for any $t \in [T]$ and $\boldsymbol{x} \in \mathcal{D}_t$, λ satisfies the following inequality,

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} - \boldsymbol{x}_t^* \rangle \leq \lambda \cdot \max_{j \in [i-1]} \langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle, \quad i \in [m]. \tag{10}$$

λ quantifies the trade-off among different objectives, which is smaller than a commonly used trade-off parameter called global trade-off (Miettinen 1999, Definition 2.8.5). Further discussion about λ is provided in the subsequent section.

Equipped with this new introduced parameter λ, we design a novel Lexicographic Ordered Arm Filter (LOAF), as detailed in Algorithm 2. At the start, LOAF initializes the candidate arm set $\mathcal{D}^0$ with the input arm set $\mathcal{D} \subseteq \mathcal{D}_t$. Then, LOAF sequentially refines the candidate arm set from the first objective to the last objective through the intersection of the scaled confidence intervals.

Specifically, for the i-th objective, LOAF first selects the arm with highest upper confidence bound from $\mathcal{D}^{i-1}$, i.e., $\hat{\boldsymbol{x}}_t^i = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{D}^{i-1}} \hat{y}_t^i(\boldsymbol{x}) + w_t(\boldsymbol{x})$. Then, LOAF retains the arms whose upper confidence bound of the i-th objective are not far from that of $\hat{\boldsymbol{x}}_t^i$, i.e.,

$$\begin{aligned} \mathcal{D}^i = \big\{ \boldsymbol{x} \in \mathcal{D}^{i-1} | \hat{y}_t^i(\boldsymbol{x}) + w_t(\boldsymbol{x}) \geq \hat{y}_t^i(\hat{\boldsymbol{x}}_t^i) + w_t(\hat{\boldsymbol{x}}_t^i) \\ -(2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1}) \cdot W \big\} \end{aligned} \tag{11}$$

where W is the upper bound of the confidence term for arms in $\mathcal{D}$, i.e., $\max_{\boldsymbol{x} \in \mathcal{D}} w_t(\boldsymbol{x}) \leq W$. The scaling parameter $2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1}$ is carefully designed to avoid losing the optimal arm $\boldsymbol{x}_t^*$. After performing this filtering process for all m objectives, LOAF outputs the arm set $\mathcal{D}^m$. The theoretical guarantee for LOAF is outlined as follows.

Proposition 1 *For Algorithm 2, if $\boldsymbol{x}_t^* \in \mathcal{D}$ and $w_t(\boldsymbol{x}) \leq W$ for any $\boldsymbol{x} \in \mathcal{D}$, then with probability at least $1-\delta$, $\boldsymbol{x}_t^* \in \mathcal{D}^m$ and for any $\boldsymbol{x} \in \mathcal{D}^m$,*

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle \leq 4(1 + \lambda + \lambda^2 + \cdots + \lambda^{i-1}) \cdot W, i \in [m].$$

Proposition 1 states that if the input arm set $\mathcal{D}$ has the optimal arm $\boldsymbol{x}_t^*$, LOAF does not lose $\boldsymbol{x}_t^*$ during the filtering process, which means the indicator function of $\widehat{R}^i(T)$ can be removed. Meanwhile, Proposition 1 provides an upper bound on the reward gap between $\boldsymbol{x}_t^*$ and other arms in $\mathcal{D}^m$.

Multiple Trade-offs Approach Taking LOAF as the arm filter, we have removed the indicator function of the metric $\widehat{R}^i(T)$. Now, our focus shifts towards employing LOAF effectively to improve the regret bound in Theorem 1.

As expressed in Proposition 1, LOAF requires the confidence terms are smaller than W. Similar to the utilization of the chain relationship in STLO, one possible approach is to divide the decision-making process into two cases:

(a) If the confidence term of all arms in $\mathcal{D}_t$ is smaller than W, apply LOAF to filter out the promising arms and randomly play an arm from these arms;

(b) If there exists an arm in $\mathcal{D}_t$ with confidence term larger than W, play this arm to reduce its uncertainty.

Proposition 1 indicates that in case (a), a smaller value of W yields a better output arm set $\mathcal{D}^m$. However, a small value of W leads an increased number of trials in case (b), which is pure exploration and results in a large instantaneous regret. Consequently, setting W as small as possible is undesirable.

To address this dilemma, we propose a multiple trade-offs approach, which divides the decision-making process at each round into multiple stages and executes a more refined trade-off between exploration and exploitation as the stages progress. The new-designed algorithm is called MTLO, whose details are shown in Algorithm 3.

At each round t, MTLO first calculates the estimators $\hat{\boldsymbol{\theta}}_t^i$ for each objective $i \in [m]$ by Eq. (6). It then computes the estimated rewards and confidence interval width for each

Algorithm 3: Multiple Trade-offs under Lexicographic Ordering (MTLO)

Require: trade-off parameter λ, time horizon T

for $t = 1, 2, \ldots, T$ **do**
  Compute the estimators $\hat{\boldsymbol{\theta}}_t^i, i \in [m]$ by Eq. (6)
  Compute the estimated rewards and confidence terms for all arms in $\mathcal{D}_t$ by Eq. (7)
  Initialize $s = 1, \mathcal{D}_{t,1} = \mathcal{D}_t$
  repeat
    if $w_t(\boldsymbol{x}) \leq 1/\sqrt{T}$ for any $\boldsymbol{x} \in \mathcal{D}_{t,s}$ **then**
      Invoke the Algorithm 2 to filter the promising arms $\mathcal{D}_{t,T} = \text{LOAF}\ (\mathcal{D}_{t,s}, 1/\sqrt{T})$
      Randomly play an arm $\boldsymbol{x}_t \in \mathcal{D}_{t,T}$ and observe $[y_t^1, y_t^2, \ldots, y_t^m]$
    else if $w_t(\boldsymbol{x}_t) > 2^{-s}$ for some $\boldsymbol{x}_t \in \mathcal{D}_{t,s}$ **then**
      Play the arm $\boldsymbol{x}_t$ and observe $[y_t^1, y_t^2, \ldots, y_t^m]$
    else
      Invoke the Algorithm 2 to filter the promising arms $\mathcal{D}_{t,s+1} = \text{LOAF}\ (\mathcal{D}_{t,s}, 2^{-s})$
      Update $s = s + 1$
    end if
  until an arm $\boldsymbol{x}_t$ is played.
  Update $X_{t+1} = [\boldsymbol{x}_\tau^\top]_{\tau \in [t]}, Y_{t+1}^i = [y_\tau^i]_{\tau \in [t]}, i \in [m]$
end for

arm in $\mathcal{D}_t$, using the formula (7). Next, MTLO initiates a repeat-until loop to iteratively refine the promising arms, starting with $\mathcal{D}_{t,1} = \mathcal{D}_t$ and $s = 1$. At each stage s, MTLO **first** checks if the confidence terms for all arms in $\mathcal{D}_{t,s}$ are less than or equal to $1/\sqrt{T}$. If this is the case, MTLO invokes LOAF with input arm set $\mathcal{D}_{t,s}$ and maximum confidence width $1/\sqrt{T}$, obtaining the promising arm set $\mathcal{D}_{t,T}$. Then, MTLO randomly plays an arm $\boldsymbol{x}_t \in \mathcal{D}_{t,T}$ and records its rewards. **Alternatively**, if the confidence term of some arm in $\mathcal{D}_{t,s}$ exceeds 2^{-s}, MTLO plays this arm for exploration and records its rewards. **Lastly**, if all arms in $\mathcal{D}_{t,s}$ have confidence terms less than or equal to 2^{-s}, MTLO invokes LOAF with input arm set $\mathcal{D}_{t,s}$ and maximum confidence width 2^{-s}, which filters out a promising arm set $\mathcal{D}_{t,s+1}$. MTLO then increases the stage index to $s + 1$ and proceeds into the next stage. As the stage goes, the maximum confidence width $W = 2^{-s}$ decreases, leading to a more refined filtering procedure in LOAF.

Let $S = \lfloor \ln T \rfloor$. Since $2^{-S} < 1/\sqrt{T}$, the repeat-until loop terminates before the index s reaches S. After observing the rewards, MTLO updates the contextual information matrix and historical reward vectors to prepare for the next round. We have the following theorem for MTLO.

Theorem 2 *Suppose that* (2) *and* (3) *hold. If MTLO is run with* λ *satisfying* (10)*, then with probability at least* $1 - \delta$*, for any objective* $i \in [m]$*, its regret satisfies*

$$R^i(T) \leq 4(1+\lambda+\lambda^2+\cdots+\lambda^{i-1}) \cdot \left(\sqrt{T} + 10\gamma_T \ln T \sqrt{dT}\right)$$

where $\gamma_T = R\sqrt{d \ln(2mT/\delta)} + U$.

Remark 2 Theorem 2 states that MTLO achieves a regret bound of $\widetilde{O}((1 + \lambda^{i-1})d\sqrt{T})$ for the i-th objective, which is almost optimal in terms of d and T (Dani, Hayes, and Kakade 2008). For the first objective, MTLO achieves a regret bound of $\widetilde{O}(d\sqrt{T})$, which aligns with the single objective SLB algorithms (Abbasi-yadkori, Pál, and Szepesvári 2011), and the increased regret for the subsequent objectives is the cost of simultaneously optimizing multiple objectives.

Remark 3 If the arms are finite, by leveraging the technique of Chu et al. (2011), we can easily obtain a regret bound of $\widetilde{O}((1+\lambda^{i-1})\sqrt{dT})$, which improves upon the existing regret bound $\widetilde{O}((KT)^{\frac{2}{3}})$ (Hüyük and Tekin 2021). Furthermore, we adopt the general regret (4), which is more accurate than the priority-based regret (1) (Hüyük and Tekin 2021).

Experiments

In this section, we present the empirical performance of our algorithms. We adopt PF-LEX (Hüyük and Tekin 2021) and OFUL (Abbasi-yadkori, Pál, and Szepesvári 2011) as baselines, where PF-LEX is designed for lexicographic MOMAB, and OFUL is a single objective SLB algorithm.

To compare with PF-LEX, we fix the arm sets as $\mathcal{D}_t = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_K\} \subseteq \mathbb{R}^d$ for any $t \geq 1$. Both the arm number K and feature dimension d are set as 10, which ensures that PF-LEX and our algorithms encounter the same number of unknown parameters. For $k \in [K]$, the arm vector $\tilde{\boldsymbol{x}}_k$ is set as the standard basis in $\mathbb{R}^d$, whose k-th element is 1 and all other elements are 0. The number of objectives is set as $m = 3$. We denote the inherent vectors as $\boldsymbol{\theta}_*^i = [\theta_*^i(1), \theta_*^i(2), \ldots, \theta_*^i(10)], i \in [3]$. The elements of $\boldsymbol{\theta}_*^1$, $\boldsymbol{\theta}_*^2$ and $\boldsymbol{\theta}_*^3$ are specified as $\theta_*^1(k) = 1 - \min_{p \in \{0.2, 0.8, 1.6\}} |0.2k - p|$, $\theta_*^2(k) = 1 - \lambda \times \min_{p \in \{0.6, 1.4\}} |0.2k - p|$ and $\theta_*^3(k) = 1 - \lambda \times |0.2k - 1.5|$, $k \in [10]$. Here, we set λ to be 0.2 and 2 to explore the performance of all algorithms across different problems.

For any arm $\tilde{\boldsymbol{x}}_k \in \mathcal{D}_t$ and $i \in [m]$, its expected reward is $\langle \boldsymbol{\theta}_*^i, \tilde{\boldsymbol{x}}_k \rangle = \theta_*^i(k)$ as $\tilde{\boldsymbol{x}}_k$ is the standard basis of $\mathbb{R}^d$. According to the design of $\theta_*^1(k)$, $\{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_4, \tilde{\boldsymbol{x}}_8\}$ are the optimal arms in the first objective. Similarly, the optimal arms for both the first and second objectives are $\{\tilde{\boldsymbol{x}}_4, \tilde{\boldsymbol{x}}_8\}$. Therefore, all three objectives has to be considered to determine the lexicographic optimal arm $\{\tilde{\boldsymbol{x}}_8\}$. After playing the arm $\boldsymbol{x}_t$, the reward of the i-th objective is sampled from a Gaussian distribution with mean $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t \rangle$ and variance 0.1. To reduce the randomness, each algorithm is run ten times, and we report the average regret. We set $T = 5 \times 10^4$ and $\delta = 0.01$. The exploration parameter ϵ for STLO and PF-LEX is set to $d^{\frac{4}{3}}(KT)^{-\frac{1}{3}}$ and $(KT)^{-\frac{1}{3}}$, respectively, which are theoretically optimal. To accelerate the convergence, we scale the confidence terms of all algorithms by 0.1, which is a common practice in bandits (Li et al. 2012; Lu et al. 2019).

Figure 2 displays the general regret curves for the first and third objectives. Subfigure (a) shows the results for $\lambda = 0.2$, while subfigure (b) presents the results for $\lambda = 2$. In Figure 2(a), OFUL demonstrates the best performance in the first objective, but shows linear regret in the third objective. The regret curves of MTLO flatten for both the first and second objectives, indicating successful identification of

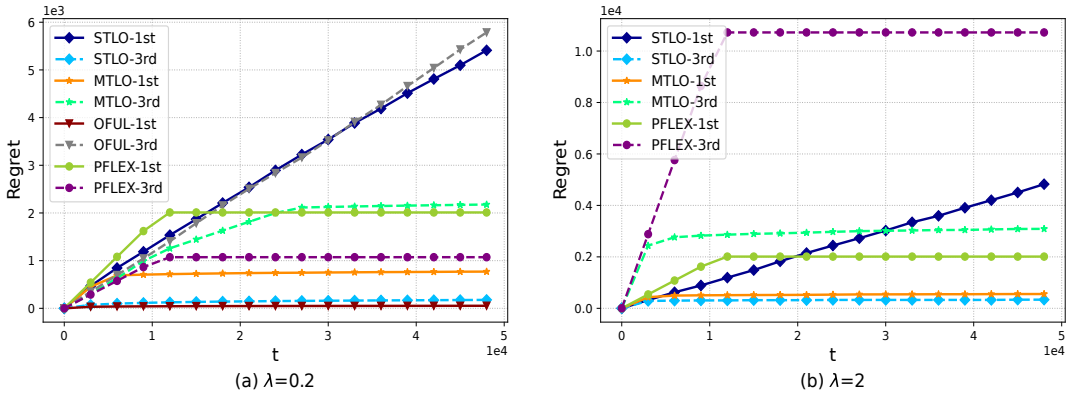

Figure 2: Comparison of our algorithms versus OFUL and PF-LEX.

the lexicographic optimal arm. An interesting observation is that while the regret curve of STLO's third objective flattens, its first objective experiences linear regret. This discrepancy is due to the fact that the theoretically optimal ϵ for STLO is 0.27 in our experiments, while the expected reward gaps for the first and second objectives are 0.2 and 0.04, respectively, indicating that most arms in $\mathcal{D}_t$ are ϵ-chained during the filtering operation. Consequently, the filtered arm set $\mathcal{D}_t^2$ contains almost all arms. Step 12 of STLO selects the arm with the highest confidence bound on the third objective, resulting in small regret on the third objective. PF-LEX avoids this issue as its theoretically optimal ϵ is 0.012 in our experiments, allowing successful arm filtering based on the chain relation. Although the regret curves of PF-LEX eventually flatten, it consumes too many trials for exploration, leading to higher regrets than MTLO in the first objective.

In Figure 2(b), there are two notable differences compared to Figure 2(a). First, the regret curve for the third objective of MTLO flattens more quickly in Figure 2(b). This is because, based on the $\{\tilde{\boldsymbol{x}}_k\}_{k\in[K]}$ and $\{\boldsymbol{\theta}_*^i\}_{i\in[3]}$ in our experiments, $\lambda = 2$ leads to a larger reward gap, aiding the identification of the optimal arm. Second, we do not display the regret curves for OFUL in Figure 2(b) due to the large regret associated with its third objective. Including these curves would obscure the plots of other algorithms. We provide the regret curves of OFUL in Appendix E.

Further Discussion

This section provides a further discussion about the necessity of introducing a new parameter to depict the relationship among objectives in lexicographic bandit problem.

In the single objective bandit problem, the arm number K or the dimension d is sufficient to reflect the difficulty of finding the optimal arm. This is because K or d is the number of unknown parameters. However, in the lexicographic bandit problem, the high-dimensional reward space brings more unknown parameters, and the lexicographic order restricts the location of the optimal reward in the reward space. The difficulty brought by the added unknown parameters can be reflected by the objective number m, while the restriction on the optimal reward requires a new parameter to depict. Thus, we introduce the trade-off parameter λ in Eq. (10).

Moreover, employing the information about trade-offs is common in various contexts (Athanassopoulos and Podinovski 1997; Nowak and Trzaskalik 2022). For instance, Keeney (2002) employs the "value trade-offs" to describe the extent to which a learner is willing to sacrifice one objective in pursuit of a specific gain in another objective. Similarly, the term "trading-off" is employed to depict a scenario where a learner, seeking incline in one criterion, must accept a decline in another criterion (Ruiz et al. 2019). The global trade-off is a ratio determining how much the value of one criterion will rise per unit decrease in another criterion when transitioning between decisions (Kaliszewski 2000). In our context, the trade-off parameter λ is a ratio indicating how much the value of the i-th objective will increase per unit decrease in the preceding $i-1$ objectives when transitioning between the optimal arm $\boldsymbol{x}_*$ and other arms. Meanwhile, λ is not confined to a specific value, and any upper bound of λ is allowed in Algorithm 3. An example of such estimates is the judgement "1 tonne of sulphur dioxide (SO2) emissions is at most 10 times as harmful as one tonne of carbon oxide (CO) emissions", where $\lambda = 10$ (Podinovski 1999).

Conclusion and Future Work

We investigated lexicographic online learning in MOSLB and extended the metric of lexicographic bandits from the priority-based regret (1) to the more accurate regret (4). We presented two algorithms: STLO and MTLO. STLO is a linear bandit adaptation of PF-LEX (Hüyük and Tekin 2021), and we reduced PF-LEX's computational complexity by the ϵ-chain relation. MTLO settles the infinite-armed setting and improves the regret bound to $\widetilde{O}((1+\lambda^{i-1})d\sqrt{T})$ for the objective $i \in [m]$, which is almost optimal in terms of d and T. MTLO's novelties include a new arm filter and a multiple trade-off approach. These techniques can be easily adapted to other bandit models, such as finite-armed SLB (Chu et al. 2011), generalized linear bandits (Jun et al. 2017) and unimodal bandits (Yu and Mannor 2011).

Currently, MTLO's regret bound exhibits an exponential relationship between λ and i. Future work will focus on reducing this and constructing a matching lower bound.

Acknowledgments

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China [GRF Project No. CityU 11215622].

References

Abbasi-yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems 24*, 2312–2320.

Agrawal, R. 1995. The Continuum-Armed Bandit Problem. *SIAM Journal on Control and Optimization*, 33(6): 1926–1951.

Alieva, A.; Cutkosky, A.; and Das, A. 2021. Robust Pure Exploration in Linear Bandits with Limited Budget. In *Proceedings of the 38th International Conference on Machine Learning*, 187–195.

Athanassopoulos, A. D.; and Podinovski, V. V. 1997. Dominance and Potential Optimality in Multiple Criteria Decision Analysis with Imprecise Information. *The Journal of the Operational Research Society*, 48(2): 142–150.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2009. Exploration–exploitation Tradeoff Using Variance Estimates in Multi-armed Bandits. *Theoretical Computer Science*, 410(19): 1876–1902.

Auer, P. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3(11): 397–422.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2–3): 235–256.

Auer, P.; Chiang, C.-K.; Ortner, R.; and Drugan, M. 2016. Pareto Front Identification from Stochastic Bandit Feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 939–947.

Azizi, M.; Kveton, B.; and Ghavamzadeh, M. 2022. Fixed-Budget Best-Arm Identification in Structured Bandits. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2798–2804.

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122.

Bubeck, S.; Dekel, O.; Koren, T.; and Peres, Y. 2015. Bandit Convex Optimization: $\sqrt{T}$ Regret in One Dimension. In *Proceedings of the 28th Conference on Learning Theory*, 266–278.

Cai, X.-Q.; Zhang, P.; Zhao, L.; Jiang, B.; Sugiyama, M.; and Llorens, A. J. 2023. Distributional Pareto-Optimal Multi-Objective Reinforcement Learning. In *Advances in Neural Information Processing Systems 36*, 15593–15613.

Cheng, J.; Xue, B.; Yi, J.; and Zhang, Q. 2024. Hierarchize Pareto Dominance in Multi-Objective Stochastic Linear Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 11489–11497.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 208–214.

Combes, R.; and Proutiere, A. 2014. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, 521–529.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the 21st Annual Conference on Learning*, 355–366.

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

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

Feng, Y.; Huang, Z.; and Wang, T. 2022. Lipschitz Bandits with Batched Feedback. In *Advances in Neural Information Processing Systems 35*, 19836–19848.

He, J.; Zhou, D.; Zhang, T.; and Gu, Q. 2022. Nearly Optimal Algorithms for Linear Contextual Bandits with Adversarial Corruptions. In *Advances in Neural Information Processing Systems 35*, 34614–34625.

Hu, J.; Chen, X.; Jin, C.; Li, L.; and Wang, L. 2021. Near-Optimal Representation Learning for Linear Bandits and Linear RL. In *Proceedings of the 38th International Conference on Machine Learning*, 4349–4358.

Huang, J.; Dai, Y.; and Huang, L. 2022. Adaptive Best-of-Both-Worlds Algorithm for Heavy-Tailed Multi-Armed Bandits. In *Proceedings of the 39th International Conference on Machine Learning*, 9173–9200.

Huang, X.; Wang, P.; Cheng, X.; Zhou, D.; Geng, Q.; and Yang, R. 2019. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2702–2719.

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

Jin, T.; Xu, P.; Xiao, X.; and Anandkumar, A. 2022. Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-Armed Bandits. In *Advances in Neural Information Processing Systems 35*, 38475–38487.

Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, 325–333.

Jun, K.-S.; Bhargava, A.; Nowak, R.; and Willett, R. 2017. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems 30*, 99–109.

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

Keeney, R. L. 2002. Common Mistakes in Making Value Trade-Offs. *Operations Research*, 50(6): 935–945.

Kone, C.; Kaufmann, E.; and Richert, L. 2024. Bandit Pareto Set Identification: the Fixed Budget Setting. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2548–2556.

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

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, L.; Chu, W.; Langford, J.; Moon, T.; and Wang, X. 2012. 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*.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

Li, W.; Barik, A.; and Honorio, J. 2022. A Simple Unified Framework for High Dimensional Bandit Problems. In *Proceedings of the 39th International Conference on Machine Learning*, 12619–12655.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019. Multi-Objective Generalized Linear Bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3080–3086.

Masoudian, S.; Zimmert, J.; and Seldin, Y. 2022. A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback. In *Advances in Neural Information Processing Systems 35*, 11752–11762.

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

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

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

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

Rodriguez, M.; Posse, C.; and Zhang, E. 2012. Multiple Objective Optimization in Recommender Systems. In *Proceedings of the 6th ACM Conference on Recommender Systems*, 11–18.

Ruiz, A. B.; Ruiz, F.; Miettinen, K.; Delgado-Antequera, L.; and Ojalehto, V. 2019. NAUTILUS Navigator: free search interactive multiobjective optimization without trading-off. *Journal of Global Optimization*, 74(2): 213–231.

Turgay, E.; Oner, D.; and Tekin, C. 2018. Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 1673–1681.

Van Moffaert, K.; Van Vaerenbergh, K.; Vrancx, P.; and Nowe, A. 2014. Multi-objective $\mathcal{X}$-Armed bandits. In *2014 International Joint Conference on Neural Networks*, 2331–2338.

Xu, M.; and Klabjan, D. 2023. Pareto Regret Analyses in Multi-objective Multi-armed Bandit. In *Proceedings of the 40th International Conference on International Conference on Machine Learning*, 38499–38517.

Xue, B.; Cheng, J.; Liu, F.; Wang, Y.; and Zhang, Q. 2024. Multiobjective Lipschitz Bandits under Lexicographic Ordering. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 16238–16246.

Xue, B.; Wang, G.; Wang, Y.; and Zhang, L. 2020. Nearly Optimal Regret for Stochastic Linear Bandits with Heavy-Tailed Payoffs. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2936–2942.

Xue, B.; Wang, Y.; Wan, Y.; Yi, J.; and Zhang, L. 2023. Efficient Algorithms for Generalized Linear Bandits with Heavy-tailed Rewards. In *Advances in Neural Information Processing Systems 36*, 70880–70891.

Yang, S.; Ren, T.; Shakkottai, S.; Price, E.; Dhillon, I. S.; and Sanghavi, S. 2022. Linear Bandit Algorithms with Sublinear Time Complexity. In *Proceedings of the 39th International Conference on Machine Learning*, 25241–25260.

Yu, J. Y.; and Mannor, S. 2011. Unimodal bandits. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 41–48.

Yu, X.; Shao, H.; Lyu, M. R.; and King, I. 2018. Pure Exploration of Multi-Armed Bandits with Heavy-Tailed Payoffs. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 937–946.

Zhang, L.; Yang, T.; Jin, R.; Xiao, Y.; and Zhou, Z. 2016. Online Stochastic Linear Optimization under One-bit Feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, 392–401.

Zhu, Y.; and Mineiro, P. 2022. Contextual Bandits with Smooth Regret: Efficient Learning in Continuous Action Spaces. In *Proceedings of the 39th International Conference on Machine Learning*, 27574–27590.