Hierarchize Pareto Dominance in Multi-Objective Stochastic Linear Bandits

Ji Cheng[1,2], **Bo Xue**[1,2], **Jiaxiang Yi**[3], **Qingfu Zhang**[1,2*]

[1]Department of Computer Science, City University of Hong Kong
[2]The City University of Hong Kong Shenzhen Research Institute
[3]Department of Material Engineering, Delft University of Technology
{J.Cheng, boxue4-c}@my.cityu.edu.hk, J.Yi@tudelft.nl, qingfu.zhang@cityu.edu.hk

Abstract

Multi-objective Stochastic Linear bandit (MOSLB) plays a critical role in the sequential decision-making paradigm, however, most existing methods focus on the Pareto dominance among different objectives without considering any priority. In this paper, we study bandit algorithms under mixed Pareto-lexicographic orders, which can reflect decision makers' preferences. We adopt the Grossone approach to deal with these orders and develop the notion of Pareto-lexicographic optimality to evaluate the learners' performance. Our work represents a first attempt to address these important and realistic orders in bandit algorithms. To design algorithms under these orders, the upper confidence bound (UCB) policy and the prior free lexicographical filter are adapted to approximate the optimal arms at each round. Moreover, the framework of the algorithms involves two stages in pursuit of the balance between exploration and exploitation. Theoretical analysis as well as numerical experiments demonstrate the effectiveness of our algorithms.

Introduction

Multi-armed bandit (MAB) model is a general paradigm for sequential decision-making problems where at each round the player chooses an action from multiple arms and then obtains a reward from the environment. The goal of MAB model is to maximize the total cumulative reward of the chosen arms over T rounds. The MAB model involves various application (Bouneffouf and Rish 2019), e.g., recommendation system (Mary, Gaudel, and Preux 2015; Gutowski et al. 2021), clinical trials (Villar, Bowden, and Wason 2015; Durand et al. 2018), etc.

However, not all problems only have an exclusive metric (objective). A more general case is that the player faces more than one feedback when places an action, which inherently complicates the decision-making process. The preference of different objectives is a critical issue among the multi-objective multi-armed bandit (MOMAB) problems. Generally, users would like to maximize all the objectives simultaneously in the context that all the objectives are incommensurable. In this scenario, different arms are compared based on the rewards with Pareto order, and those arms with optimal rewards are called Pareto optimal arms, in which the optimality is judged by Pareto dominance. To gauge the performance of a MOMAB algorithm, Pareto regret was proposed (Drugan and Nowe 2013) based on the gap between the selected arm and the Pareto optimal arms. The goal of the algorithm is to select arms that judiciously minimize the Pareto regret according to historical observations. For example, agents in online advertising system have to consider the click-through rate for exposure as well as the click conversion for revenue simultaneously (Rodriguez, Posse, and Zhang 2012).

In some scenarios, the objectives are ranked lexicographically (Ehrgott 2005), which means that the first objective holds absolute priority over the second one, which in turn has higher precedence over the third one, etc. The bandit algorithm for these problems aims to select arms that yield high rewards for all the objectives, however, the improvement of the rewards in the low-priority objectives is satisfied only if the objectives that have higher priority do not sacrifice (Hüyük and Tekin 2021). Real-world applications include learning optimal routing for wireless sensor networks (Shah-Mansouri, Mohsenian-Rad, and Wong 2008), delivering doses of radiation to target volumes or to normal tissues in intensity modulated radiation treatment (IMRT) (Jee, McShan, and Fraass 2007), etc.

In this work, we focus on a more general order, mixed Pareto-lexicographic order, which hierarchizes the Pareto dominance to consider Pareto order and lexicographical order simultaneously. To the best of our knowledge, this is the first work to analyze multi-objective stochastic linear bandit (MOSLB) problems under the mixed Pareto-lexicographic order. Note that by setting the precedence among objectives, the mixed order can specialize in the Pareto order or lexicographical order. The contributions of this work are summarized as follows.

- We formulated the MOSLB framework under two kinds of mixed Pareto-lexicographic orders, specifically mixed Pareto-lexicographic order under priority chains (MPL-PC) and mixed Pareto-lexicographic order under priority levels (MPL-PL).
- Aimed at the two MOSLB problems, we defined the optimality justification and the regret evaluation based on gross-scalars which properly denotes the infinite and in-

*Corresponding author.

finitesimal quantities, respectively.

- Finally, we developed two MOSLB algorithms under the two mixed Pareto-lexicographic orders, respectively. Theoretical analysis of the regret shows that the algorithms are Hannan consistent, and the effectiveness of the algorithms is verified through numerical experiments.

Remainder of the paper is organized as follows. We review the related work in Section 2. The bandit framework under the novel mixed orders is introduced in Section 3, followed by the proposed learning algorithms in Section 4. Section 5 demonstrates the study of numerical experiments. The conclusion and future work are summarized in the last part.

Related Work

In this section, we give a discussion on related work with stochastic bandits, multi-objective bandits, and the Grossone methodology.

Stochastic Bandits

In 1985, Lai, Robbins et al. (1985) developed a stochastic MAB algorithm which is upper bounded by $O(K \log T)$ and provided a matching lower bound. Auer (2002) proposed the SupLinRel algorithm with respect to linear model which yielded regret bound of $\widetilde{O}(\sqrt{dT})$, where a sophisticated device is developed to decouple reward dependence. The confidence region approach was then proposed in (Dani, Hayes, and Kakade 2008) to derive the upper confidence bound of the expected reward, with which the algorithm results in the regret bound of $\widetilde{O}(d\sqrt{T})$. Later, Abbasi-Yadkori, Pál, and Szepesvári (2011) improved the regret bound by a logarithmic factor. The upper confidence bound (UCB) based technique has been thoroughly explored in stochastic bandit paradigm over decades (Auer 2002; Abbasi-Yadkori, Pál, and Szepesvári 2011; Bubeck et al. 2015; Zhang et al. 2016; Xue et al. 2020; Hu et al. 2021; Alieva, Cutkosky, and Das 2021; He et al. 2022; Li, Barik, and Honorio 2022; Masoudian, Zimmert, and Seldin 2022; Jin et al. 2022).

Another piece of the research assumes the relation between the contextual information and the reward can be modeled as a Lipschitz function (Kleinberg, Slivkins, and Upfal 2008; Lu et al. 2019b). Lu, Pál, and Pál (2010) presented a Query-Ad-Clustering algorithm with its regret upper bounded by $\tilde{O}\left(T^{1-1/(2+d_c)}\right)$ and lower bounded by $\Omega\left(T^{1-1/(2+d_p)}\right)$. Slivkins (2011) developed a method, contextual zooming, which achieves near-optimal regret bounds $\tilde{O}\left(T^{1-1/(2+d_z)}\right)$ for the problems with exponentially or infinitely large strategy sets. Recently, Feng, Huang, and Wang (2022) developed a novel algorithm called Batched Lipschitz Narrowing (BLiN) which can optimally solve the problem of Lipschitz bandit problems with regret rate $\widetilde{O}(T^{\frac{d_z+1}{d_z+2}})$ using $O(\log \log T)$ batches.

Multi-objective Bandits

To tackle the problems with multi-dimensional rewards, Drugan and Nowe (2013) first introduced the MOMAB with Pareto order and developed the algorithms which took the upper bound $O(K \log T)$ of the Pareto regret. An algorithm for multi-objective contextual bandit problems where one objective dominates the other was proposed in (Tekin and Turgay 2018) and achieves $\widetilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ on both their developed 2D regret and Pareto regret. Turgay, Oner, and Tekin (2018) then studied the bandit model with contextual information, where the expected reward satisfied the Lipschitz condition. Later, Lu et al. (2019a) improved MOMAB to tackle non-linear rewards by cooperating with generalized linear bandit model and obtained a Pareto regret bound $\widetilde{O}(d\sqrt{T})$. Hüyük and Tekin (2021) first analyzed the MOMAB under lexicographic ordering and developed a priority-based regret to assess the bandit algorithm under this environment. Their developed algorithm obtained a suboptimal upper bound $\widetilde{O}\left(K^{\frac{2}{3}}T^{\frac{2}{3}}\right)$ for the priority-based regret. Recently, Xu and Klabjan (2023) presented new algorithms and analyses for adversarial MOMAB, providing insights into the formulation of Pareto regrets and their applications. Besides, researchers also focused on the identification of Pareto optimal arms within limited budget (Van Moffaert et al. 2014; Auer et al. 2016; Kone, Kaufmann, and Richert 2023; Kim, Iyengar, and Zeevi 2023).

Grossone Methodology

Grossone methodology (GM) introduced a novel numeral system proposed and developed by Sergeyev (2017) to represent finite, infinite, and infinitesimal numbers with Grossone base, ①. Due to the fact that the axioms of infinite and infinitesimal quantities complement the axioms of the real numbers perfectly, four basic operations as well as the comparison operator are well defined for the Grossone base. Moreover, the standard properties (commutative, associative, existence of inverse, etc.) also work for the numerals of GM. The method has already successfully applied to optimization (Cococcioni, Pappalardo, and Sergeyev 2018; Lai, Fiaschi, and Cococcioni 2020; Lai et al. 2021), differential equations (Sergeyev 2013), game theory (Fiaschi and Cococcioni 2021), and so forth.

Instead of using ∞, a numeral associated with infinite or infinitesimal quantities can be denoted by the GM as follows,

$$c = c_{p_m}①^{p_m} + \cdots + c_{p_0}①^{p_0} + c_{p_{-1}}①^{p_{-1}} + \cdots + c_{p_{-k}}①^{p_{-k}}$$

where $m, k \in \mathbb{N}$. The exponents p_i and the digits c_{p_i} are called *gross*-powers and *gross*-digits, respectively. A quantity with *gross*-powers being zero represents a real number, while infinite and infinitesimal quantities correspond to positive and negative *gross*-powers, respectively. For example, $7.8①^2 + 3①^0 - 2.1①^{-1}$ is a *gross*-scalar with one infinite element, one finite element, and an infinitesimal element correspondingly.

Let $f_1, \ldots, f_m$ denote the m objectives, the lexicographically optimization problems can be reformulated by using *gross*-scalar as,

$$\min \ f_1(\mathbf{x}) + ①^{-1} f_2(\mathbf{x}) + \cdots + ①^{1-m} f_m(\mathbf{x}). \qquad (1)$$

The priority relation is rooted within the *gross*-powers: the higher the power, the larger the priority. The most impor-

tant objective $f_1(\mathbf{x})$ is indeed associated with the exponent 0.

Problem Description

In this section, we first review the problem setting of MOSLB and the learning goals under Pareto order and lexicographic order, then we develop the MOSLB framework under two more generally applicable orders, mixed Pareto-lexicographic orders.

In a T-round MOSLB problem, the agent observes the context of K arms $\mathcal{X}_t = \{\mathbf{x}_{t,a} \in \mathbb{R}^d \mid a \in [K]\}$ at round $t \in T$. Once the learner selects the arm $a_t \in [K]$, it receives a reward vector $\boldsymbol{y}(\mathbf{x}_{t,a_t}) = [y_{t,1}, y_{t,2}, \ldots, y_{t,m}] \in \mathbb{R}^m$ with $y_{t,m} = y_m(\mathbf{x}_{t,a_t})$ for brevity. In stochastic linear bandits, the rewards are supposed to be random variables with expectations

$$\mathbb{E}\left[y_i(\mathbf{x}_{t,a}) \mid \mathbf{x}_{t,a}, \mathcal{F}_{t-1}\right] = \langle \boldsymbol{\theta}_i^*, \mathbf{x}_{t,a} \rangle, \forall i \in [m], \quad (2)$$

where $\boldsymbol{\theta}_i^*$ is unknown parameter to be estimated for the i-th objective, and $\mathcal{F}_{t-1} = \{\mathbf{x}_{1,a_1}, \ldots, \mathbf{x}_{t-1,a_{t-1}}\} \cup \{y_{1,1}, y_{2,1}, \ldots, y_{t-1,1}\} \cup \ldots \cup \{y_{1,m}, y_{2,m}, \ldots, y_{t-1,m}\}$ constitutes a σ-filtration of events up to round t. We assume that the stochastic rewards are sub-Gaussian.

The goal of the learner is to minimize the accumulated regret over T rounds, where the regret for each round is generally measured through the gap between the selected arm and the optimal arms. However, due to the fact that the objectives in a multi-objective problems are conflict with each other, which means that to improve one objective has to sacrifice the other one. Pareto order and lexicographic order are two commonly used metrics to measure the preference between multiple objectives, and the detail of these two orders is introduced as follows.

Pareto Order

When the objectives are incomparable, the optimality is judged based on Pareto dominance defined as follows.

Definition 1 *(Pareto dominance). Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$ be two vectors in objective space, $\boldsymbol{u}$ is said to Pareto-dominate $\boldsymbol{v}$, denoted as $\boldsymbol{u} \succ_{par} \boldsymbol{v}$, if and only if $\forall i \in [m], u_i \geq v_i$ and $\exists j \in [m], u_j > v_j$.*

Given the Pareto dominance, the concept of Pareto suboptimality gap (PSG) was introduced in (Drugan and Nowe 2013) to measure expected loss between the chosen arm and optimal arms.

Definition 2 *(PSG). Let $\mathbf{x}$ be the context in $\mathcal{X}$, and $\boldsymbol{\mu}(\mathbf{x})$ be the vector of its expected reward. The Pareto suboptimality gap is defined as the minimal scalar $\zeta \geq 0$ such that by adding ζ to all entries of the expected reward the arm can not be Pareto dominated by any other arms. Formally,*

$$\Delta(\mathbf{x}) = \inf\left\{\zeta \in \mathbb{R}_+ \mid (\boldsymbol{\mu}(\mathbf{x}) + \zeta) \not\succ_{par} \boldsymbol{\mu}(\mathbf{x}'), \forall \mathbf{x}' \in \mathcal{X}\right\}.$$

In this way, performance of the learner can be evaluated by Pareto regret, which is given as

$$PR(T) = \sum_{t=1}^{T} \Delta(\mathbf{x}_{t,a_t}). \quad (3)$$

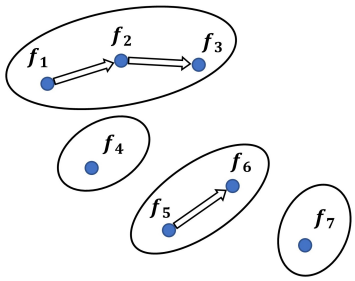

Figure 1: Illustration of MPL-PC, where each dot denotes one objective and arrows represent precedence relationship. (In the example, there are preferences among the objectives f_1, f_2, and f_3, where f_1 is indefinitely important than f_2 which matters more than f_3. The similar relation happens between f_5 and f_6. Besides, there is not any priority between the circles.)

Lexicographic Order

In consideration of the other scenario where the objectives are ranked lexicographically, lexicographic dominance can be defined as

Definition 3 *(Lexicographic dominance). Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$ be two vectors in objective space, $\boldsymbol{u}$ is said to lexicographically dominate $\boldsymbol{v}$ in the first i objectives, denoted as $\boldsymbol{u} \succ_{lex,i} \boldsymbol{v}$, if and only if $u_j > v_j$, where $j = \min\{k \leq i : u_k \neq v_k\}$.*

Given lexicographic dominance, an arm a_t^* is said to be lexicographically optimal if and only if there is no arm in the decision set that can lexicographically dominate it in all the m objectives. Hüyük and Tekin (2021) developed the priority based regret for MOSLB under lexicographic order as

$$LR_i(T) = \sum_{t=1}^{T} \left(\mu_i(\mathbf{x}_{t,a_t^*}) - \mu_i(\mathbf{x}_{t,a_t})\right) \mathbb{I}(A_{t,i}), \quad (4)$$

where $A_{t,i} = \left\{\mu_j(\mathbf{x}_{t,a_t^*}) = \mu_j(\mathbf{x}_{t,a_t}), 1 \leq j \leq i-1\right\}$ and $\mathbb{I}(\cdot)$ is the indicator function.

Mixed Pareto-lexicographic Orders

In this subsection, We introduce two realistic and general orders for multi-objective bandit problems and develop the optimality gap and algorithm regret for MOSLB.

MPL-PC Based on users' preference, there may exist lexicographical relationship among part of the objectives, for example in Fig 1. In this setting, we can partition the objective space $\mathcal{Y}$ to c subspace $\mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_c$, where the objectives in $\mathcal{Y}_i \subset \mathbb{R}^{m_i}$ are ranked lexicographically and $\sum_{i=1}^{c} m_i = m$. We have no priority between different subspace $\mathcal{Y}_i$, which means the objectives from different subspace have to treat equivalently. Followed (Lai, Fiaschi, and Cococcioni 2020), the order under this scenario is called mixed Pareto-lexicographic order under priority chains, MPL-PC.

Based on the GM, we can denote the rewards at an arm a by a vector of *gross*-scalar $\boldsymbol{y}(\mathbf{x}_{t,a}) =$

$[y_{(1)}(\mathbf{x}_{t,a}), y_{(2)}(\mathbf{x}_{t,a}), \ldots, y_{(c)}(\mathbf{x}_{t,a})]^\top$ with $y_{(i)}(\mathbf{x}_{t,a}) = y_{(i),1}(\mathbf{x}_{t,a}) + y_{(i),2}(\mathbf{x}_{t,a})①^{-1} + \cdots + y_{(i),m_i}(\mathbf{x}_{t,a})①^{1-m_i}$, where $y_{(i),1}$ is the most important objective in $\mathcal{Y}_i$, followed by $y_{(i),2}$, etc. Inspired by this representation, we can define the optimality condition under this order as follows.

Definition 4 *(MPL-PC optimality). A vector with gross-scalar entries $\boldsymbol{u} = [u_{(1)}, u_{(2)}, \ldots, u_{(c)}]$ is said to MPL-PC optimal if there is no other $\boldsymbol{v}$ such that $v_{(j)} \geq u_{(j)}$, $\forall j \in [c]$ and $\boldsymbol{u} \neq \boldsymbol{v}$.*

To gauge the performance of bandit algorithms, we need to define a regret quantity which measures the gap in metrics between a_t and a_t^*. Due to the properties of GM representation, the gap between the chosen arm and the optimal arms can be defined by changing the real number to *gross*-scalar as follows.

Definition 5 *(MPL-PC suboptimality gap). Let $\mathbf{x}$ be the arm pulled by the learner. The suboptimality gap w.r.t. MPL-PC is defined as the minimum gross-scalar with non-negative digits $c = c_0 + c_1①^{-1} + c_2①^{-2} + \cdots$, $c_i \geq 0$ that by adding to the scalar to each chain of the expected reward the arm becomes Pareto-lexicographic optimal.*

$$\widetilde{\Delta}(\mathbf{x}) = \inf \{c \mid \boldsymbol{\mu}(\mathbf{x}_{t,a_t}) + c \not\prec_{par} \boldsymbol{\mu}(\mathbf{x}'), \forall \mathbf{x}' \in \mathcal{X}\}.$$

This definition is similar to that under Pareto order except that the gap is measured by a *gross*-scalar, which inherits the lexicographic relationship in it. In this way, we can evaluate the learner's performance under MPL-PC by

$$PCR(T) = \sum_{t=1}^{T} \widetilde{\Delta}(\mathbf{x}_{t,a_t}). \tag{5}$$

MPL-PL We also consider another scenario where a group of the objectives is infinitely more important than the second group which in turn is comparably more important than the third one, etc. The objectives among the same group are incommensurable, for example in Fig. 2. This order was termed as mixed Pareto-lexicographic under priority levels, MPL-PL. Interested readers can refer to (Lai et al. 2021) for real-world applications under this scenario. Supposed that we have l groups of objectives whose preferences are ordered lexicographically. By partitioning the objective space $\mathcal{Y}$ to $\mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \mathcal{Y}_l$ where $\mathcal{Y}_i \subset \mathbb{R}^{m_i}$ and $\sum_{i=1}^{l} m_i = m$, we denote the rewards as $\boldsymbol{y}(\mathbf{x}_{t,a}) = [\boldsymbol{y}_1(\mathbf{x}_{t,a})^\top, \ldots, \boldsymbol{y}_l(\mathbf{x}_{t,a})^\top]^\top$. Then the optimality under this order can be formulated as follows.

Definition 6 *(MPL-PL optimality). A vector $\boldsymbol{u} = [\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, \ldots, \boldsymbol{u}_l^\top]^\top \in \mathcal{Y}$ is said to be optimal in the first k priority level if there is no $\boldsymbol{v} \in \mathcal{Y}$ such that $\boldsymbol{v}_j \succ_{par} \boldsymbol{u}_j$, $\forall j \in [k]$.*

We leverage the *gross*-scalar with *gross*-digits being Pareto suboptimality gap for each priority levels to evaluate the performance of the chosen arms. Let $\Delta_i(\mathbf{x}_{t,a_t})$ be the PSG at round t for the i-th priority level, then the regret for a bandit algorithm can be formulated as

$$PLR(T) = \sum_{t=1}^{T} \sum_{i=1}^{l} ①^{i-1} \Delta_i(\mathbf{x}_{t,a_t}) \cdot \mathbb{I}(\tilde{A}_{t,i}), \tag{6}$$

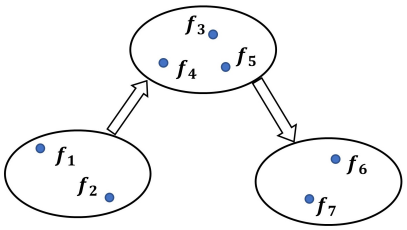

Figure 2: Illustration of MPL-PL. (The first two objectives f_1 and f_2 are preferred than the next three objectives, moreover, the objectives in the circle are incomparable.)

where $\tilde{A}_{t,i} = \{\Delta_j(\mathbf{x}_{t,a_t}) = 0, 1 \leq j \leq i-1\}$.

Remark 1 *Noted that these two orders are a more general order compared to order that only involves Pareto or lexicographical dominance, specifically, the order may specialize in Pareto dominance with no chain exist in MPL-PC and lexicographic order with each group in MPL-PL contains an exclusive objective.*

The Learning Algorithms

In this part, we develop the stochastic bandit algorithms with MPL-PC and MPL-PL, respectively. Without loss of generality, we assume that $\|\mathbf{x}_{t,a}\| \leq 1$ for $a \in [K]$ and $\|\boldsymbol{\theta}_i^*\| \leq 1$, $i \in [m]$. Throughout the paper, $\|\mathbf{x}\|$ is the l-2 norm of vector $\mathbf{x} \in \mathbb{R}^d$, and the induced norm of $\mathbf{x}$ by a positive definite matrix $V \in \mathbb{R}^{d \times d}$ is $\|\mathbf{x}\|_V = \sqrt{\mathbf{x}^\top V \mathbf{x}}$.

MOSLB-PC

We first propose a MOSLB algorithm for problems with the MPL-PC order, MOSLB-PC, which is outlined in Algorithm 1. The algorithm starts with pure exploration stage followed by focused exploitation process. At initial rounds, the uncertainty of the estimated reward may severely influence the decision process, therefore, we consider to lower the width of confidence interval of each arm down to the threshold ϵ. Then in exploitation process, the algorithm focuses on determining the optimal arms for each priority chain separately.

Let $X_t = [\mathbf{x}_{1,a_1}, \ldots, \mathbf{x}_{t-1,a_{t-1}}]^\top \in \mathbb{R}^{(t-1)\times d}$ be the matrix of past decisions, and $Y_{t,i} = [y_i(\mathbf{x}_{1,a_1}), \ldots, y_i(\mathbf{x}_{t-1,a_{t-1}})]^\top \in \mathbb{R}^{(t-1)\times 1}$ be historical rewards of the i-th objective. At each round, the algorithm needs to estimate the parameter $\boldsymbol{\theta}_i^*$ for i-th objective by l^2-regularized least squares estimate with regularization parameter $\lambda = 1$ as,

$$\hat{\boldsymbol{\theta}}_{t,i} = (X_t^\top X_t + I)^{-1} X_t^\top Y_{t,i}. \tag{7}$$

According to the theorem of self-normalized bound (Abbasi-Yadkori, Pál, and Szepesvári 2011) for martingales, the parameter $\boldsymbol{\theta}_i^*$ lies in the an ellipsoid centered at $\hat{\boldsymbol{\theta}}_{t,i}$ with probability at least $1-\delta$, therefore, a confidence set $\mathcal{C}_{t,i}$ can be constructed as,

$$\mathcal{C}_{t,i} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t,i}\|_{V_t} \leq \gamma_t\right\}, \tag{8}$$

Algorithm 1: MOSLB-PC

Input: Time horizon T, and confidence level $\delta \in (0,1)$

Initialize $V_1 = I_d$, $\hat{\boldsymbol{\theta}}_{1,i} = \mathbf{0},\ i \in [m]$
for $t = 1, 2, \ldots, T$ **do**
  Observe each arm's context $\mathbf{x}_{t,a}, a \in [K]$
  Evaluate $\hat{y}_{t,i}(\mathbf{x}_{t,a}), w_t(\mathbf{x}_{t,a})$ for $i \in [m], a \in [K]$
  if $w_t(\mathbf{x}_{t,a}) > \epsilon$ for some $a \in [K]$ **then**
    Play the arm a_t selected randomly from those arms and observe the reward
  else
    Compute $\boldsymbol{u}_t(\mathbf{x}_{t,a}), \boldsymbol{l}_t(\mathbf{x}_{t,a})$ based on Eq. 10
    for $j = 1, 2, \ldots, c$ **do**
      Obtain suboptimal arm set $\mathcal{A}^*_{t,(j)}$ for each chain based on the PFLF algorithm
    end for
    Compute optimal arm set $\mathcal{A}^*_t = \bigcup_{i\in[c]} \mathcal{A}^*_{t,(i)}$
    Play an arm a_t from $\mathcal{A}^*_t$ uniformly at random and observe the reward
  end if
  Update $V_{t+1} = V_t + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$, X_{t+1}, and $Y_{t+1,i}$
  Update the estimators $\hat{\boldsymbol{\theta}}_{t+1,i} = V_{t+1}^{-1} X_{t+1}^\top Y_{t+1,i}$
end for

Algorithm 2: Prior Free Lexicographical filter (PFLF)

Input: Objective space $\mathcal{Y}_c \subset \mathbb{R}^{m_c}$, upper and lower confidence bounds for each arm $a \in [K]$
Output: $\mathcal{A}^*_{(c)} = \mathcal{A}^*_{(c),m_c}$

Initialize $\mathcal{A}^*_{(c),0} = [K]$
for $i = 1, 2, \ldots, m_c$ **do**
  Evaluate $\hat{a}^*_i = \operatorname{argmax}_{a\in\mathcal{A}^*_{(c),i-1}} u_{t,i}(\mathbf{x}_{t,a})$
  Filter the arms $\mathcal{A}^*_{(c),i} = \{a \in \mathcal{A}^*_{(c),i-1} : aC_i\hat{a}^*_i\}$
end for

where, $\gamma_t = R\sqrt{d\log\left(\frac{m(1+t)}{\delta}\right)} + 1$ and $V_t = V_{t-1} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$, $V_0 = \lambda I_d$.

Based on the the principle of "optimism in face of uncertainty", we can evaluate the upper confidence bound of the expected reward for any arm a by

$$u_{t,i}(\mathbf{x}) = \max_{\boldsymbol{\theta}\in\mathcal{C}_{t,i}} \boldsymbol{\theta}^\top \mathbf{x},\ i \in [m]. \tag{9}$$

Moreover, by algebraic manipulation, the optimization can be solved with the closed form:

$$u_{t,i}(\mathbf{x}) = \hat{y}_{t,i}(\mathbf{x}) + w_t(\mathbf{x}),\ i \in [m] \tag{10}$$

where $\hat{y}_{t,i}(\mathbf{x}) = \langle\hat{\boldsymbol{\theta}}_{t,i}, \mathbf{x}\rangle$ is the estimated expected reward, $w_t(\mathbf{x}) = \gamma_t\|\mathbf{x}\|_{V_t^{-1}}$ denotes the width of confidence level. Similarly, the lower confidence bound can be calculated as $l_{t,i}(\mathbf{x}) = \hat{y}_{t,i}(\mathbf{x}) - w_t(\mathbf{x})$. The upper confidence bounds can be written as the form $\boldsymbol{u}_t(\mathbf{x}) = [u_{t,(1)}(\mathbf{x}), \ldots, u_{t,(c)}(\mathbf{x})]^\top$ based on the predefined representation.

A prior free filter algorithm, as shown in Algorithm 2 is adapted to deal with the lexicographically ranked objectives in exploitation stage. In the algorithm, the arm with the highest upper confidence bound for the first objective in the chain is determined as the best arm. Considered the uncertainty, the two arms a and a' are said to be linked in objective i if $[l_{t,i}(\mathbf{x}_{t,a}), u_{t,i}(\mathbf{x}_{t,a})] \cap [l_{t,i}(\mathbf{x}_{t,a'}), u_{t,i}(\mathbf{x}_{t,a'})] \neq \emptyset$. Then those arms which are in the same component of the transitive closure of the linked relation in objective i with the best arm $\hat{a}^*_i$, written as $aC_i\hat{a}^*_i$, are selected as the sub-optimal arms. After the PFLF algorithm assesses the suboptimal arm set for each priority chain, the overall suboptimal arms are formed as the union of the suboptimal set for each chain due to incomparable relationship among different chains. Finally, the algorithm selects an arm uniformly at random from the optimal arm set $\mathcal{A}^*_t$ and updates the parameters based on the reward received from the environment. The following theorem establishes the theoretical guarantees for the MOSLB-PC algorithm.

Theorem 1 *Assume the arm set is finite, i.e., $K < \infty$, $\forall t \in [T]$, and the maximum length of the priority chains is $|c|_{max} = \max_{i\in[c]} m_i$. If the MOSLB-PC algorithm runs with $\delta \in (0,1)$ and $\epsilon > 0$, then with probability at least $1-\delta$,*

$$PCR(T) \leq \sum_{i=1}^{|c|_{max}} ①^{i-1} \left(100\epsilon^{-2}d\gamma_T^2 \log T + 2TK\epsilon\right)$$

where, $\gamma_T = R\sqrt{d\log(m(1+T)/\delta)} + 1$.

Remark: Theorem 1 implies that the algorithm involves the upper-bounded regret of $\widetilde{O}\left((dKT)^{2/3}\right)$ without any prior knowledge by setting the parameter $\epsilon = d^{2/3}(KT)^{-1/3}$. The element of the regret matches that of the existing algorithm for bandit under lexicographic order (Hüyük and Tekin 2021). Notably, the regret correlates with the number of objectives in the longest priority chain.

MOSLB-PL

Now we focus on bandit problems with MPL-PL, and the developed algorithm, MOSLB-PL, is presented in Algorithm 3. The process similarly involves the exploration and exploitation stages. In the exploitation stage, we obtain the optimal arms level by level by considering the upper confidence bound of each arm. Followed (Lu et al. 2019a), the empirical Pareto optimality is adapted to determine the optimal arms at each priority level, where those arms that their upper confidence bound cannot be Pareto dominated by that of the other are recognized. Naturally, the optimal arms for the last level form the final optimal arm set at this round, and we select an arm from it randomly. The performance of MOSLB-PL can be guaranteed through the following theorem, and the detailed proofs can be found in the Appendix.

Theorem 2 *Assume the arm set is finite. If the MOSLB-PL algorithm runs with $\delta \in (0,1)$ and $\epsilon > 0$, then with probability at least $1-\delta$,*

$$PLR(T) \leq \sum_{i=1}^{l} ①^{i-1} \left(100\epsilon^{-2}d\gamma_T^2 \log T + 2T\epsilon\right).$$

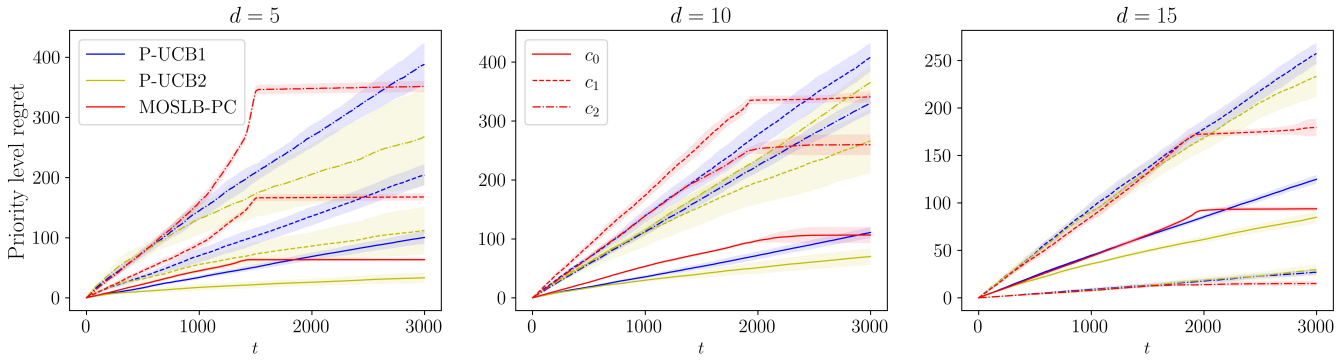

Figure 3: Pareto chain regret of the compared methods for MPL-PC with two chains: $c_0 + c_1①^{-1} + c_2①^{-2}$.

Remark: Theorems 2 shows that expected loss of MOSLB-PL is upper bounded by $\widetilde{O}\left((dT)^{2/3}\right)$ with the parameter $\epsilon = d^{2/3}(T)^{-1/3}$. Besides, the length of the *gross*-scalar for the regret is determined by the number of priority levels.

Experiments

In this section, we conducted numerical experiments to verify the performance of our proposed bandit algorithms. The method developed (Drugan and Nowe 2013; Lu et al. 2019a) was adopted as the baseline, denoted as P-UCB in our setting. P-UCB considers the objectives in Pareto order and compares the arms based on the upper confidence bound of the estimated rewards.

MPL-PC

We first consider the MPL-PC scenario, and the comparative methods are listed as follows.

- P-UCB1: This method discards the lexicographic relationship within priority chains and treats all the objectives equivalently in Pareto dominance.
- P-UCB2: This method only considers the first objectives in each priority chain, optimizing c objectives in Pareto order.

We first experimented in the environment with $m = 5$ objectives and the priority chains represented by $\{(1, 2), (3, 4, 5)\}$, which means that the first chain contains two objectives in lexicographic order while the second chain has three objectives. Three settings, the context's dimension d are picked from $\{5, 10, 15\}$, were investigated, and the unknown coefficients $\boldsymbol{\theta}_i^*$ are sampled uniformly from the unit ball. We generated $5d$ arms uniformly from the centered unit ball.

Since the algorithms involve randomness, we carried out 10 trials with round $T = 3000$ and reported the outcomes in Fig. 3, where the lines represent average performance among ten trials and the shadow area shows the variance. The regret is denoted as *gross*-scalar $c_0 + c_1①^{-1} + c_2①^{-2}$ which is impractical to draw this scalar in Cartesian coordinates. Therefore, we plotted the iterative process of its *gross*-digits, where c_0 presents the regret of the most important objective in each priority chain followed by c_1 and c_2. Implementation code can be accessed via our webpage[1].

It can be seen from the results that P-UCB2 performs well for the regret of the first priority as it only considers the most important objective. P-UCB1 performs the worst and attains almost linear regret which means that treating all objectives in Pareto order do not work. On the contrary, the proposed MOSLB-PC converges efficiently after the stage of exploration on all the three hierarchies, which demonstrates that the proposed method can well address the decision-making problems with MPL-PC order.

Algorithm 3: MOSLB-PL

Input: Time horizon T, and confidence level $\delta \in (0, 1)$

Initialize $V_1 = I_d$, $\hat{\boldsymbol{\theta}}_{1,i} = \mathbf{0}$, $i \in [m]$
for $t = 1, 2, \ldots, T$ **do**
  Observe each arm's context $\mathbf{x}_{t,a}, a \in [K]$
  Evaluate $\hat{y}_{t,i}(\mathbf{x}_{t,a}), w_t(\mathbf{x}_{t,a})$ for $i \in [m], a \in [K]$
  if $w_t(\mathbf{x}_{t,a}) > \epsilon$ for some $a \in [K]$ **then**
    Play the arm a_t selected randomly from those arms and observe the reward
  else
    Compute the upper confidence bound $\boldsymbol{u}_t(\mathbf{x}_{t,a})$
    Initialize $\mathcal{A}^*_{t,0} = [K]$
    for $j = 1, 2, \ldots, l$ **do**
      Obtain suboptimal set $\mathcal{A}^*_{t,j} = \{a \in \mathcal{A}^*_{t,j-1} \mid \boldsymbol{u}_{t,j}(\mathbf{x}_{t,a'}) \not\succ_{\text{par}} \boldsymbol{u}_{t,j}(\mathbf{x}_{t,a}), \forall a' \in \mathcal{A}^*_{t,j-1}\}$
    end for
    Play an arm a_t from $\mathcal{A}^*_{t,l}$ uniformly at random and observe the reward
  end if
  Update $V_{t+1} = V_t + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$, X_{t+1}, and $Y_{t+1,i}$
  Update the estimators $\hat{\boldsymbol{\theta}}_{t+1,i} = V_{t+1}^{-1} X_{t+1}^\top Y_{t+1,i}$
end for

MPL-PL

To verify the effectiveness of the MOSLB-PL algorithm, we compared it with the same baseline methods of the last ex-

[1] https://github.com/jicheng9617/moslb

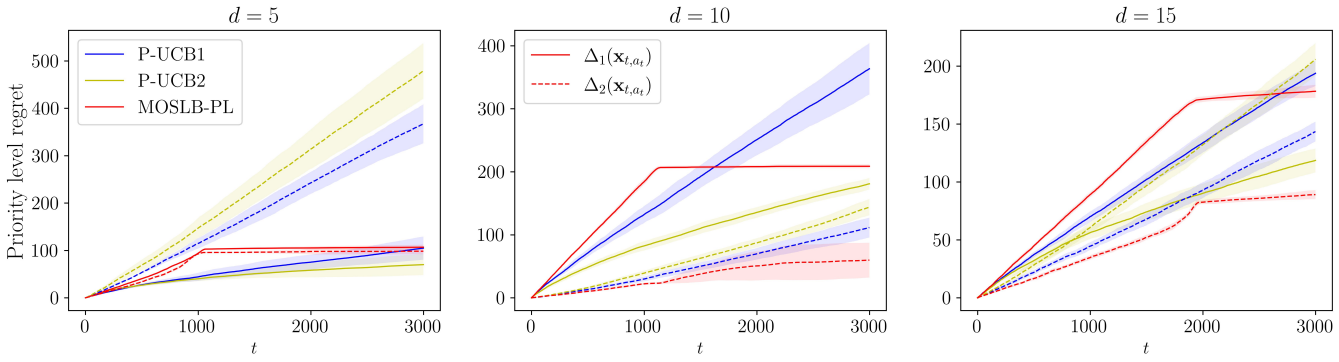

Figure 4: Pareto level regret of the compared methods for MPL-PL with two levels: $\Delta_1(\mathbf{x}_{t,a_t}) + \Delta_2(\mathbf{x}_{t,a_t})①^{-1}$.

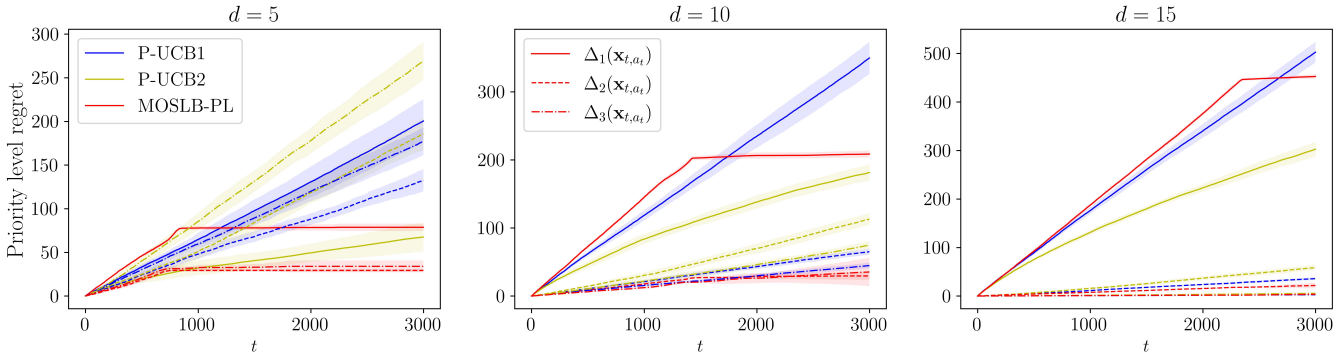

Figure 5: Pareto level regret of the compared methods for MPL-PL: $\Delta_1(\mathbf{x}_{t,a_t}) + \Delta_2(\mathbf{x}_{t,a_t})①^{-1} + \Delta_3(\mathbf{x}_{t,a_t})①^{-2}$.

periment except that the P-UCB2 method only considers the objectives in the first priority level.

In this experiment, environments are set the same as that in previous one, and we first consider the priority level represented by $\{(1, 2, 3), (4, 5)\}$ with $m = 5$ objectives. The notation means that the five objectives are segmented into two groups, where the first one consists of three objectives and is infinitely important than the second group which contains the last two objectives. The results are shown in Fig. 4, where we can see that the proposed algorithm outperforms the baselines w.r.t. the mean and the variance of the regret. It should be noted that the Pareto level regret involves the indicator function, therefore, the regret in the second level accumulates only when the algorithm chooses the arms that have no regret in the first level.

Furthermore, to test the performance of the MOSLB-PL algorithm when facing more priority levels, we performed the experiment with MPL-PL problems represented by $\{(1, 2, 3), (4, 5, 6), (7, 8, 9, 10)\}$. The results are shown in the Fig. 5. It can be observed that the proposed algorithm can efficiently choose the optimal arms for multi-objective problems under MPL-PL. On the contrary, the P-UCB1 method, which treats all objectives jointly, fails in the experiment, and P-UCB2 performs well in the first level but accumulates regret at second and third levels.

Conclusions

In this paper, we studied and developed the MOSLB framework under two realistic orders with users' preferences, MPL-PC and MPL-PL. Based on the principle of optimism in face of uncertainty and the balance of exploration and exploitation, we developed two UCB-type MOSLB algorithms for the two mixed orders, respectively. To measure the performance of the algorithms, we adopted the Grossone methodology to represent Pareo-lexicographic optimality, through which the precedence relationship can be inherited in the representation of the scalar. The element of the regret denoted by *gross*-scalar of the proposed algorithms is upper bounded by $\widetilde{O}\left((dT)^{2/3}\right)$, which is in line with that for MOSLB under pure lexicographic order. The performance is analyzed theoretically and verified through numerical experiments compared with the baseline methods.

In future, we will further investigate the applications of the proposed algorithms in real-world multi-objective sequential decision-making problems. Besides, the optimal arm identification approach may be meaningful but scarce for the applications of the mixed Pareto-lexicographic orders.

Acknowledgments

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Adminis-

trative Region, China [GRF Project No: CityU 11215622] and by Natural Science Foundation of China (Project No: 62276223).

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. In *Advances in neural information processing systems*, volume 24, 2312–2320.

Alieva, A.; Cutkosky, A.; and Das, A. 2021. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, 187–195.

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

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.

Bouneffouf, D.; and Rish, I. 2019. A Survey on Practical Applications of Multi-Armed and Contextual Bandits. arXiv:1904.10040.

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.

Cococcioni, M.; Pappalardo, M.; and Sergeyev, Y. D. 2018. Lexicographic multi-objective linear programming using grossone methodology: Theory and algorithm. *Applied Mathematics and Computation*, 318: 298–311.

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 (IJCNN)*, 1–8.

Durand, A.; Achilleos, C.; Iacovides, D.; Strati, K.; Mitsis, G. D.; and Pineau, J. 2018. Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, 67–82.

Ehrgott, M. 2005. *Multicriteria optimization*, volume 491. Springer Science & Business Media.

Feng, Y.; Huang, z.; and Wang, T. 2022. Lipschitz Bandits with Batched Feedback. 19836–19848.

Fiaschi, L.; and Cococcioni, M. 2021. Non-Archimedean game theory: A numerical approach. *Applied Mathematics and Computation*, 409: 125356.

Gutowski, N.; Amghar, T.; Camp, O.; and Chhel, F. 2021. Gorthaur-EXP3: Bandit-based selection from a portfolio of recommendation algorithms balancing the accuracy-diversity dilemma. *Information Sciences*, 546: 378–396.

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*, volume 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 *International Conference on Machine Learning*, 4349–4358.

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

Jee, K.-W.; McShan, D. L.; and Fraass, B. A. 2007. Lexicographic ordering: intuitive multicriteria optimization for IMRT. *Physics in Medicine & Biology*, 52(7): 1845.

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*, 38475–38487.

Kim, W.; Iyengar, G.; and Zeevi, A. 2023. Pareto Front Identification with Regret Minimization.

Kleinberg, R.; Slivkins, A.; and Upfal, E. 2008. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 681–690.

Kone, C.; Kaufmann, E.; and Richert, L. 2023. Adaptive Algorithms for Relaxed Pareto Set Identification.

Lai, L.; Fiaschi, L.; and Cococcioni, M. 2020. Solving mixed Pareto-Lexicographic multi-objective optimization problems: The case of priority chains. *Swarm and Evolutionary Computation*, 55: 100687.

Lai, L.; Fiaschi, L.; Cococcioni, M.; and Deb, K. 2021. Solving mixed pareto-lexicographic multiobjective optimization problems: the case of priority levels. *IEEE Transactions on Evolutionary Computation*, 25(5): 971–985.

Lai, T. L.; Robbins, H.; et al. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22.

Li, W.; Barik, A.; and Honorio, J. 2022. A simple unified framework for high dimensional bandit problems. In *International Conference on Machine Learning*, 12619–12655.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019a. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3080–3086.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019b. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, 4154–4163.

Lu, T.; Pál, D.; and Pál, M. 2010. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, 485–492.

Mary, J.; Gaudel, R.; and Preux, P. 2015. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data*, 325–336.

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*, volume 35, 11752–11762.

Rodriguez, M.; Posse, C.; and Zhang, E. 2012. Multiple objective optimization in recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*, 11–18.

Sergeyev, Y. 2013. Solving Ordinary Differential Equations by Working with Infinitesimals Numerically on the Infinity Computer. *Applied Mathematics and Computation*, 219(22): 10668–10681.

Sergeyev, Y. D. 2017. Numerical infinities and infinitesimals: Methodology, applications, and repercussions on two Hilbert problems. *EMS Surveys in Mathematical Sciences*, 4(2): 219–320.

Shah-Mansouri, V.; Mohsenian-Rad, A.-H.; and Wong, V. W. 2008. Lexicographically optimal routing for wireless sensor networks with multiple sinks. *IEEE Transactions on Vehicular Technology*, 58(3): 1490–1500.

Slivkins, A. 2011. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, 679–702.

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

Turgay, E.; Oner, D.; and Tekin, C. 2018. Multi-objective Contextual Bandit Problem with Similarity Information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 1673–1681.

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

Villar, S. S.; Bowden, J.; and Wason, J. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science*, 30(2): 199.

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

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 Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2936–2942.

Zhang, L.; Yang, T.; Jin, R.; Xiao, Y.; and Zhou, Z.-H. 2016. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, 392–401.