Safe Multi-Objective Linear Bandits with Hierarchical Preferences

Bo Xue[1,2], **Mengxia He**[3], **Yilu Liu**[1,2], **Ji Cheng**[1,2], **Zhe Zhao**[1] and **Qingfu Zhang**[1,2,*]
[1]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[2]The City University of Hong Kong Shenzhen Research Institute, Shenzhen, China
[3]Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong, China
{boxue4-c, yilu.liu, J.Cheng}@my.cityu.edu.hk, mxhe@eee.hku.hk, zz4543@mail.ustc.edu.cn,
qingfu.zhang@cityu.edu.hk

Abstract

Multi-objective bandits with hierarchical preferences and safety constraints is central to many real-world decision-making tasks such as healthcare treatment planning and safe autonomous control, where multiple objectives must be optimized according to their priorities while ensuring safety requirements are satisfied. In this paper, we study a multi-objective stochastic linear bandit framework that incorporates *hierarchical preferences* together with *safety constraints*, requiring the learner to remain competitive with respect to a known baseline policy. We consider two practically motivated safety models: **(i)** *cumulative constraints*, which require the cumulative performance to exceed the baseline, and **(ii)** *stage-wise constraints*, which impose this requirement at each time step. We propose two algorithms, LexUCB-C and LexTS-S, designed for the cumulative and stage-wise settings, respectively. We establish regret bounds showing that both algorithms achieve performance comparable to existing single-objective safe linear bandit methods, while simultaneously optimizing multiple objectives. In addition to theoretical guarantees, we develop an experimental framework that captures the interaction between hierarchical preferences and safety constraints. Experiments on synthetic and real-world datasets demonstrate the effectiveness of the proposed methods.

1 Introduction

Multi-armed bandits (MAB) provide a foundational framework for modeling sequential decision-making under uncertainty [Lai and Robbins, 1985; Auer *et al.*, 2002; Abbasi-yadkori *et al.*, 2011; Lattimore and Szepesvári, 2020]. In a standard MAB problem, a learner interacts with an environment over T rounds. At each round $t \in [T]$, the learner selects an arm from a finite or structured decision set and receives a stochastic reward corresponding to the chosen arm. The goal of the learner is to maximize the cumulative expected reward. A key challenge in this problem is the exploration-exploitation trade-off: the learner must balance exploring less-known arms to gather information and exploiting currently known arms that believed to yield high rewards.

To address this trade-off, two widely used strategies have been developed: Upper Confidence Bound (UCB) and Thompson Sampling (TS). UCB algorithms construct confidence intervals for each arm's expected reward and select the arm with the highest upper bound [Auer, 2002; Dani *et al.*, 2008; Magureanu *et al.*, 2014; Xue *et al.*, 2020, 2023]. This strategy favors arms with either high empirical rewards or high uncertainty, thereby balancing exploration and exploitation. In contrast, TS adopts a Bayesian perspective by maintaining a posterior distribution over the arm parameters and selecting arms according to their probability of being optimal, thus encouraging exploration through randomization [Thompson, 1933; Chapelle and Li, 2011; Xu *et al.*, 2023].

While the classical MAB focuses on a single objective [Bubeck and Cesa-Bianchi, 2012], many real-world problems involve multiple, often conflicting objectives. For example, recommender systems [Rodriguez *et al.*, 2012] must balance user satisfaction, content diversity, and long-term engagement, while autonomous driving [Zhang *et al.*, 2023] requires trade-offs among safety, passenger comfort, and energy efficiency. To capture such multi-criteria decision-making scenarios, the multi-objective multi-armed bandit (MOMAB) framework has been proposed [Drugan and Nowe, 2013], in which the learner receives a vector-valued reward at each round, with each component representing a distinct objective.

However, the MOMAB framework treats each arm independently and thus fails to exploit contextual or structural information often available in real-world applications, such as product categories, prices, or user profiles in recommender systems [Li *et al.*, 2010]. To address this limitation, the multi-objective stochastic linear bandit (MOSLB) model was later introduced [Lu *et al.*, 2019; Xue *et al.*, 2025], where each arm is associated with a feature vector, and the expected reward for each objective is assumed to be a linear function of this vector. This assumption enables the algorithm to generalize across arms: by leveraging the shared linear structure, the learner can use feedback from one arm to estimate the rewards of other arms with similar features.

Most existing methods in the multi-objective bandit literature rely on scalarization [Drugan and Nowe, 2013; Wanigasekara *et al.*, 2019] or Pareto-based criteria [Auer *et al.*,

*Qingfu Zhang is the corresponding author.

2016; **?**], which either collapse the multi-objective problem into a single-objective one by imposing fixed preference weights, or seek to identify Pareto-optimal solutions without explicitly modeling priority among objectives. However, these approaches may fall short in applications where objectives are inherently prioritized. For example, in clinical treatment planning, ensuring safety or efficacy is often paramount, while secondary factors such as cost or comfort are considered only after primary goals are satisfied. To accommodate such structured preferences, lexicographic ordering has been proposed [Ehrgott, 2005; Skalse *et al.*, 2022; Abernethy *et al.*, 2024], wherein objectives are ranked by importance, and reward vectors are compared based on the first dimension where they differ. Incorporating lexicographic preferences into the bandit learning framework leads to the lexicographic multi-objective bandit problem [Tekin, 2019; Hüyük and Tekin, 2021], where the higher-priority objectives must be optimized before lower-priority ones.[1]

Despite recent progress in lexicographic bandit optimization [Hüyük and Tekin, 2021; Xue *et al.*, 2024; **?**; **?**], existing methods often neglect safety constraints, which are essential in many real-world applications. For example, in high-stakes environments such as healthcare and autonomous systems, deploying exploratory policies without performance guarantees may lead to unacceptable outcomes. This motivates the study of conservative (or safe) bandits [Wu *et al.*, 2016; Katariya *et al.*, 2019; Kazerouni *et al.*, 2017; Moradipari *et al.*, 2020; Lin *et al.*, 2022], where the learner must ensure that its performance does not fall significantly below that of a known baseline policy. ***While conservative constraints have been well-studied in single-objective settings, whether comparable guarantees can be attained under multi-objective formulations remains an open problem.***

In this paper, we introduce the framework of safe MOSLB with hierarchical preferences, which combines lexicographic preferences with two types of safety constraints: cumulative constraints and stage-wise constraints. This is the first work to address multi-objective bandits under safety constraints, and our main contributions are summarized as follows:

- Under cumulative constraints, we present LexUCB-C, a UCB-based algorithm that achieves a regret bound of $\widetilde{O}(W^i(w) \cdot d\sqrt{T} + d^2 \cdot \alpha^{-2})$ for each objective $i \in [m]$ where $W^i(w) = 1+w+\cdots+w^{i-1}$, w is a weight parameter reflecting the trade-off among conflicting objectives (introduced in Assumption 4), d is the dimension, T is the time horizon, and α is the safety threshold.
- Under stage-wise constraints, we develop LexTS-S, a Thompson Sampling-based method that enforces per-round safety. LexTS-S achieves a regret bound of $\widetilde{O}(W^i(w) \cdot d^{3/2}\sqrt{T} + d \cdot \alpha^{-3})$ for each objective $i \in [m]$.
- Both algorithms achieve regret bounds comparable to those of single-objective safe bandit methods [Kazerouni *et al.*, 2017; Moradipari *et al.*, 2020], in terms of their dependence on d, T and α. Moreover, since $W^1(w) = 1$ for any trade-off value w, the performance of the highest-priority objective ($i = 1$) remains nearly unaffected when optimizing additional objectives.
- We conduct experiments on both synthetic and real-world datasets using carefully designed settings that reflect hierarchical preferences and safety requirements. The results validate our theoretical findings.

[1]Due to space limitations, a more comprehensive discussion of related work can be found in the Appendix A.

2 Problem Setting

This section formalizes the problem setting. First, we introduce the notation and the standard MOSLB model. Second, we define the lexicographic ordering and extend the conservative bandit framework to the multi-objective setting. Finally, we state the key assumptions used in our analysis.

Notation. We denote by $\boldsymbol{x} \in \mathbb{R}^d$ a feature (context) vector in a d-dimensional real space. The standard Euclidean norm of $\boldsymbol{x}$ is written as $\|\boldsymbol{x}\|$, while the Mahalanobis norm with respect to a positive definite matrix $\mathrm{V} \in \mathbb{R}^{d\times d}$ is defined as $\|\boldsymbol{x}\|_{\mathrm{V}} = (\boldsymbol{x}^\top \mathrm{V}\boldsymbol{x})^{1/2}$. For a positive integer m, we use $[m]$ to represent the index set $\{1, 2, \ldots, m\}$. The superscript $i \in [m]$ indicates quantities associated with the i-th objective.

Model. MOSLB is a T-round sequential decision-making problem. At each round $t \in [T]$, the learner observes a set of feasible arms $\mathcal{X}_t \subseteq \mathbb{R}^d$, and selects an arm $\boldsymbol{x}_t \in \mathcal{X}_t$. Upon selection, the learner receives a stochastic reward vector $\boldsymbol{y}_t = [y_t^1, y_t^2, \ldots, y_t^m] \in \mathbb{R}^m$, where m is the number of objectives. The expected reward for each objective $i \in [m]$ is assumed to follow a linear model:

$$\mathrm{E}[y_t^i] = \mu^i(\boldsymbol{x}_t) = \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t \rangle,$$

where $\boldsymbol{\theta}_*^i \in \mathbb{R}^d$ is an unknown vector for the i-th objective.

Next, we adopt the lexicographic order [Hüyük and Tekin, 2021] to compare reward vectors across multiple objectives.

Definition 1 (Lexicographic Order)**.** *Suppose $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$. $\boldsymbol{u}$ lexicographically dominates $\boldsymbol{v}$, if there exists an index $i^* \in [m]$ such that $u^j = v^j$ for all $j < i^*$, and $u^{i^*} > v^{i^*}$.*

For example, $[1, 3, 2]$ dominates $[1, 0, 8]$ with $i^* = 2$. Based on this, the optimal arm is defined as follows.

Definition 2 (Lexicographic Optimal Arm)**.** *An arm $\boldsymbol{x}_t^* \in \mathcal{X}_t$ is lexicographic optimal if its expected reward vector is not lexicographically dominated by that of any arm in $\mathcal{X}_t$.*

Given the optimal arm $\boldsymbol{x}_t^*$, we can naturally define the regret in multi-objective setting, which is the cumulative gap between the optimal arm $\boldsymbol{x}_t^*$ and the chosen arm $\boldsymbol{x}_t$, i.e.,

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

In conservative bandits, the learner is required to maintain performance that does not fall significantly below that of a given baseline policy [Kazerouni *et al.*, 2017; Moradipari *et al.*, 2020]. We extend this principle to the multi-objective setting. Specifically, at each round t, a baseline arm $\boldsymbol{x}_{b_t}$ is provided. We consider two types of safety constraints:

Definition 3 (Cumulative Constraint)**.** *For each objective $i \in [m]$, the learner's cumulative expected reward must exceed a fraction $(1-\alpha)$ of the baseline's cumulative reward:*

$$\forall t \in [T], \sum_{\tau=1}^{t} \mu^i(\boldsymbol{x}_\tau) \geq (1-\alpha) \sum_{\tau=1}^{t} \mu^i(\boldsymbol{x}_{b_\tau}).$$

Definition 4 (Stage-Wise Constraint)**.** *For each objective $i \in [m]$, the learner's expected reward at each round must exceed a fraction $(1-\alpha)$ of the baseline reward:*

$$\forall t \in [T], \mu^i(\boldsymbol{x}_t) \geq (1-\alpha)\mu^i(\boldsymbol{x}_{b_t}).$$

Finally, we introduce some common assumptions on linear bandits [Abbasi-yadkori *et al.*, 2011; Jun and Kim, 2024], conservative bandits [Kazerouni *et al.*, 2017; Moradipari *et al.*, 2020] and multi-objective bandits [Xue *et al.*, 2024].

Assumption 1. *The reward noise is σ-sub-Gaussian, i.e., for all $t \in [T]$ and $i \in [m]$,*

$$\mathrm{E}\left[e^{\beta(y_t^i - \mu^i(\boldsymbol{x}_t))}\right] \leq \exp\left(\frac{\beta^2\sigma^2}{2}\right), \forall \beta \in \mathbb{R}.$$

Assumption 2. *The inherent parameter vectors and contextual vectors are bounded, such that there exists some $B > 0$,*

$$\|\boldsymbol{\theta}_*^i\| \leq B, \forall i \in [m], \text{and } \|\boldsymbol{x}\| \leq 1, \forall \boldsymbol{x} \in \mathcal{X}_t.$$

Meanwhile, $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle \leq 1$ holds for all $i \in [m]$ and $\boldsymbol{x} \in \mathcal{X}_t$.

The sub-Gaussian assumption allows us to derive high-probability confidence bounds via standard concentration inequalities, while the boundedness of both model parameters and context vectors ensures that the expected rewards remain within a well-defined range.

Assumption 3. *Let $\Delta^i(\boldsymbol{x}_{b_t}) = \mu^i(\boldsymbol{x}_t^*) - \mu^i(\boldsymbol{x}_{b_t})$. There exist constants $0 \leq \Delta_\ell \leq \Delta_u$ and $0 \leq \mu_\ell \leq \mu_u$ such that for all $i \in [m]$ and $t \in [T]$,*

$$\Delta_\ell \leq \Delta^i(\boldsymbol{x}_{b_t}) \leq \Delta_u, \text{and } \mu_\ell \leq \mu^i(\boldsymbol{x}_{b_t}) \leq \mu_u.$$

This assumption ensures that the conservative baseline arm maintains a bounded gap from the optimal arm, which is critical for establishing safe exploration guarantees.

Assumption 4. *There exists a constant $w > 0$ such that for any $i \geq 2$ and $\boldsymbol{x} \in \mathcal{X}_t$, the trade-off between the i-th objective and higher-priority objectives is bounded as*

$$\mu^i(\boldsymbol{x}) - \mu^i(\boldsymbol{x}_t^*) \leq w \cdot \max_{j \in [i-1]} \{\mu^j(\boldsymbol{x}_t^*) - \mu^j(\boldsymbol{x})\}.$$

This structural condition captures the lexicographic preference among objectives, ensuring that, relative to the optimal arm $\boldsymbol{x}_t^*$, any improvement in a lower-priority objective cannot compensate for excessive degradation in higher-priority ones.

3 Algorithms

We develop two algorithms for lexicographic linear bandits with safety constraints. LexUCB-C addresses the cumulative constraint using the UCB principle, while LexTS-S handles the stage-wise constraint via Thompson Sampling. Together, they provide complementary approaches for safe lexicographic bandit optimization.

3.1 Lexicographic UCB for Cumulative Safety

LexUCB-C is a UCB-based algorithm for setting with cumulative safety constraints. At each round, it either selects a high-performing arm under lexicographic preferences or defaults to a conservative baseline to ensure safety. The full procedure is in Algorithm 1.

Algorithm 1 LexUCB-C

Input: $T, m, B, \alpha, \delta, \lambda, w$

Set $\mathrm{V}_1 = \lambda \mathrm{I}$, $c_1 = \sqrt{\lambda}B$, $\hat{\boldsymbol{\theta}}_1^i = \mathbf{0}$, $\boldsymbol{z}_1 = \mathbf{0}$, $\mathcal{T}_1 = \emptyset$
for $t = 1$ to T **do**
  $\boldsymbol{x}_t' \leftarrow \textsc{SelectArm}(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathrm{V}_t, \mathcal{X}_t)$ *(Alg. 2)*
  Compute the lower bounds L_t^i for $i \in [m]$ by Eq. (1)
  if Safety condition (2) holds for all objectives **then**
    Play $\boldsymbol{x}_t = \boldsymbol{x}_t'$ and observe $[y_t^1, y_t^2, \ldots, y_t^m]$
    Update $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \boldsymbol{x}_t$, $\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{t\}$
    Update $\mathrm{V}_{t+1} = V_t + \boldsymbol{x}_t \boldsymbol{x}_t^\top$
    Compute the estimators $\{\hat{\boldsymbol{\theta}}_{t+1}^i\}_{i=1}^m$ by Eq. (3)
    Compute the confidence radius c_{t+1} by Eq. (4)
  else
    Play baseline $\boldsymbol{x}_t = \boldsymbol{x}_{b_t}$ *(Conservative play)*
    Maintain $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t$, update $\mathcal{T}_{t+1}^c = \mathcal{T}_t^c \cup \{t\}$
    Keep $\mathrm{V}_{t+1} = \mathrm{V}_t$, $c_{t+1} = c_t$, $\hat{\boldsymbol{\theta}}_{t+1}^i = \hat{\boldsymbol{\theta}}_t^i, \forall i \in [m]$
  end if
end for

Initialization. At the beginning of the algorithm, several quantities are initialized. The regularized covariance matrix is set to $\mathrm{V}_1 = \lambda \mathrm{I}$, where $\lambda > 0$ ensures numerical stability in the estimation process. The confidence radius is initialized as $c_1 = \sqrt{\lambda}B$, where B is the bounded prior in Assumption 2. The initial estimators for each objective $i \in [m]$ are set to zero vectors, i.e., $\hat{\boldsymbol{\theta}}_1^i = \mathbf{0}$. The cumulative decision vector $\boldsymbol{z}_1$ is initialized to zero and will be used to accumulate the sum of selected arms that are not conservative. The set $\mathcal{T}_1$, which tracks rounds in which optimistic (i.e., non-conservative) arms are played, is initialized to be empty.

Main Loop. At each round t, LexUCB-C first calls the subroutine SELECTARM (Algorithm 2), which returns a candidate arm $\boldsymbol{x}_t'$ that respects the lexicographic preference structure. The details of this subroutine will be discussed shortly.

Given a selected candidate arm $\boldsymbol{x}_t'$, LexUCB-C evaluates its safety by computing the worst-case lower bound on the cumulative reward, including $\boldsymbol{x}_t'$ for each objective $i \in [m]$,

$$L_t^i = \min_{\boldsymbol{\theta} \in \mathcal{C}_t^i} \langle \boldsymbol{\theta}, \boldsymbol{z}_t + \boldsymbol{x}_t' \rangle, \quad \mathcal{C}_t^i = \{\boldsymbol{\theta} \| \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \leq c_t\}. \quad (1)$$

$\boldsymbol{z}_t$ denotes the cumulative action vector up to round $t-1$, and $\hat{\boldsymbol{\theta}}_t^i$ is the empirical estimate of the reward vector for objective i. The term L_t^i captures the most pessimistic cumulative reward that could be incurred under current confidence region, assuming $\boldsymbol{x}_t'$ is selected.

To ensure safety, LexUCB-C verifies whether the total worst-case reward exceeds a certain fraction of the reward obtained by always playing a fixed conservative baseline. Specifically, the arm $\boldsymbol{x}_t'$ is considered safe if the following condition holds:

$$L_t^i + \sum_{\tau \in \mathcal{T}_t^c} \mu^i(\boldsymbol{x}_{b_\tau}) \geq (1-\alpha) \sum_{\tau=1}^{t} \mu^i(\boldsymbol{x}_{b_\tau}), \forall i \in [m], \quad (2)$$

where $\mathcal{T}_t^c$ is the set of rounds up to $t-1$ in which the conservative baseline action $\boldsymbol{x}_{b_\tau}$ was played. This constraint guarantees that, in each objective, the cumulative reward does not fall below the benchmark established by the baseline policy.

Algorithm 2 SELECTARM

Input: $\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathrm{V}_t, \mathcal{X}_t$
Initialize $s = 1$, $\mathcal{X}_{t,s} = \mathcal{X}_t$
repeat
  if $\exists\, \boldsymbol{x} \in \mathcal{X}_{t,s}$ with $\|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} \leq 1/\sqrt{T}$ **then**
    $\mathcal{X}_{t,T} \leftarrow$ LEXFILTER$(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s}, \frac{1}{\sqrt{T}})$
    Return any $\boldsymbol{x} \in \mathcal{X}_{t,T}$
  else if $\exists\, \boldsymbol{x} \in \mathcal{X}_{t,s}$ with $\|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} > 2^{-s}$ **then**
    Return such $\boldsymbol{x}$
  else
    $\mathcal{X}_{t,s+1} \leftarrow$ LEXFILTER$(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s}, 2^{-s})$
    $s \leftarrow s + 1$
  end if
until an arm is selected

Algorithm 3 LEXFILTER

Input: $\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s}, C$
Initialize $\mathcal{X}_{t,s}^0 = \mathcal{X}_{t,s}$
for $i = 1$ to m **do**
  $\hat{\boldsymbol{x}}_t^i = \arg\max_{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1}} \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle$
  $\mathcal{X}_{t,s}^i = \{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1} \mid \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i - \boldsymbol{x} \rangle \leq (\sum_{j=0}^{i-1} 4w^j + 2) \cdot c_t \cdot C\}$
end for
return $\mathcal{X}_{t,s}^m$

If the safety condition is satisfied for all objectives, the candidate arm $\boldsymbol{x}_t'$ is considered safe. LexUCB-C plays $\boldsymbol{x}_t'$ and observes the reward vector $[y_t^1, \ldots, y_t^m]$. It then updates the cumulative arm vector, $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \boldsymbol{x}_t$, adds the current round to the non-conservative set, $\mathcal{T}_{t+1} = \mathcal{T}_t \cup t$, updates the covariance matrix, $\mathrm{V}_{t+1} = \mathrm{V}_t + \boldsymbol{x}_t \boldsymbol{x}_t^\top$, and recomputes the estimators $\{\hat{\boldsymbol{\theta}}_{t+1}^i\}_{i=1}^m$ via regularized least squares:

$$\hat{\boldsymbol{\theta}}_{t+1}^i = \mathrm{V}_{t+1}^{-1} \mathrm{X}_{t+1} \mathrm{Y}_{t+1}^i, \tag{3}$$

where

$$\mathrm{X}_{t+1} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t], \mathrm{Y}_{t+1}^i = [y_1^i, y_2^i, \ldots, y_t^i].$$

The confidence radius c_{t+1} is then updated accordingly, i.e.,

$$c_{t+1} = \sigma \sqrt{d \log\left(\frac{m(1 + |\mathcal{T}_{t+1}|/\lambda)}{\delta}\right)} + \sqrt{\lambda} B. \tag{4}$$

If the safety test fails for some objective, LexUCB-C plays a predefined conservative arm $\boldsymbol{x}_t = \boldsymbol{x}_{b_t}$. In this case, the cumulative vector $\boldsymbol{z}_t$, estimators, confidence radius and covariance matrix remain unchanged. The round index is added to the set of conservative rounds: $\mathcal{T}_{t+1}^c = \mathcal{T}_t^c \cup \{t\}$.

Arm Selection. To identify a candidate arm $\boldsymbol{x}_t'$ for the safety test, LexUCB-C relies on a dedicated subroutine, SELECTARM, which aims to find a lexicographically optimal arm under current uncertainty. We now describe this arm selection procedure and its internal mechanism.

SELECTARM begins with the full feasible arm set $\mathcal{X}_{t,1} = \mathcal{X}_t$ and iteratively refines it. At each iteration s, SELECTARM considers the current set $\mathcal{X}_{t,s}$, and evaluates the uncertainty term $\|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}$ for each arm $\boldsymbol{x} \in \mathcal{X}_{t,s}$. Precisely, **if** there exists an arm with sufficiently small uncertainty (Step 3), SELECTARM invokes the LEXFILTER procedure with a high-confidence threshold $C = 1/\sqrt{T}$, and randomly returns any arm from the resulting filtered set. This ensures the selection of a confident and lexicographically optimal arm. **Otherwise**, if there exists an arm with norm exceeding a coarser threshold 2^{-s} (Step 6), such an arm is directly selected. This branch encourages timely exploration when high-confidence arms are not yet available. **If neither** condition is met (Step 8), the arm set is refined by applying LEXFILTER with the threshold $C = 2^{-s}$. Note that this lexicographic criterion becomes stricter as the iteration progresses, i.e., from stage s to $s + 1$. The updated set $\mathcal{X}_{t,s+1}$ is then used in the next iteration, and such process repeats until an arm is selected.

Lexicographic Elimination. LEXFILTER performs sequential elimination across objectives following their priorities. Given the current arm set $\mathcal{X}_{t,s}$, it maintains a sequence of filtered subsets $\mathcal{X}_{t,s}^i$ for $i \in [m]$. At each level i, LEXFILTER identifies the empirically optimal arm,

$$\hat{\boldsymbol{x}}_t^i = \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1}} \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle.$$

It then eliminates any arm $\boldsymbol{x}$ from $\mathcal{X}_{t,s}^{i-1}$ if its estimated reward falls behind that of $\hat{\boldsymbol{x}}_t^i$ by more than a threshold proportional to $c_t \cdot C$,

$$\mathcal{X}_{t,s}^i = \{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1} \mid \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i - \boldsymbol{x} \rangle \leq (\sum_{j=0}^{i-1} 4w^j + 2) \cdot c_t \cdot C\}.$$

The threshold is scaled by a factor capturing the cumulative uncertainty propagated from higher-priority objectives. After m rounds, the final surviving set $\mathcal{X}_{t,s}^m$ is returned which consists of arms that remain competitive under all objectives.

In summary, LexUCB-C combines lexicographic optimization with cumulative safety monitoring. It ensures that the chosen arms respect safety constraints over time while optimizing higher-priority objectives first. We provide the following theoretical guarantee for LexUCB-C.

Theorem 1. *Suppose Assumptions 1 - 4 hold. Then, with probability at least $1 - \delta$, for any objective $i \in [m]$, the regret of LexUCB-C is bounded as*

$$R^i(T) \leq 4W^i(w) \cdot c_T \cdot \left(\sqrt{T} + 10 \log(T) \sqrt{dT}\right) + \frac{d^2 K \Delta_u}{\alpha \mu_\ell \cdot (\Delta_\ell \alpha \mu_\ell)},$$

where

$$W^i(w) = 1 + w + \cdots + w^{i-1},$$

$$c_T = \sigma \sqrt{d \log\left(\frac{m(1 + |\mathcal{T}_T|/\lambda)}{\delta}\right)} + \sqrt{\lambda} B,$$

$$K = 228(B\sqrt{\lambda} + \sigma)^2 \left[\log\left(\frac{62d\sqrt{m}(B\sqrt{\lambda} + \sigma)}{\sqrt{\delta}(\Delta_\ell + \alpha\mu_\ell)}\right)\right]^2.$$

Remark 1. Theorem 1 shows that LexUCB-C achieves a regret bound of $\widetilde{O}(W^i(w) \cdot d\sqrt{T} + d^2 \cdot \alpha^{-2})$ for any objective $i \in [m]$. The first term matches the standard regret rate of single-objective linear bandits [Kazerouni *et al.*, 2017] up to a multiplicative factor $W^i(w)$. This factor reflects the cost of optimizing multiple objectives in a lexicographic manner. Notably, $W^1(w) = 1$ implies that the regret bound on the highest-priority objective remains unaffected even optimizing other objectives simultaneously. Lower-priority objectives incur larger regret so as to maintain prioritization structure.

3.2 Lexicographic TS for Stage-Wise Safety

Under the stage-wise safety setting, we propose LexTS-S, a Thompson Sampling-based algorithm that incorporates lexicographic preferences and enforces safety at every round. The full procedure is summarized in Algorithm 4.

Initialization. At the beginning, LexTS-S set the regularized covariance matrix as $\mathrm{V}_1 = \lambda \mathrm{I}$, where the parameter $\lambda > 0$ ensures numerical stability. The initial confidence parameter is defined as $c_1 = \beta_1 = \sqrt{\lambda}B$, and the reward estimator for each objective $i \in [m]$ is initialized to zero, i.e., $\hat{\boldsymbol{\theta}}_1^i = \mathbf{0}$. Meanwhile, a mixing parameter $\rho_1 \in (0, \frac{\mu_\ell \alpha}{B+\mu_u})$ is also specified for constructing safe exploration.

Main Loop. At each round t, LexTS-S first generates a randomized parameter vector to estimate reward for each objective. Specifically, for each $i \in [m]$, a sample is drawn from a Gaussian posterior,

$$\tilde{\boldsymbol{\theta}}_t^i \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_t^i, c_t^2 \cdot \mathrm{V}_t^{-1}),$$

where the posterior mean $\hat{\boldsymbol{\theta}}_t^i$ and covariance $c_t^2 \cdot \mathrm{V}_t^{-1}$ are constructed using the same regularized least-squares estimator and covariance matrix as in LexUCB-C. This TS-based sampling introduces randomized exploration, in contrast to the optimistic UCB strategy used in LexUCB-C.

Next, LexTS-S constructs a stage-wise safety region $\mathcal{S}_t$ by intersecting the per-objective safe sets:

$$\mathcal{S}_t = \bigcap_{i=1}^{m} \mathcal{S}_t^i,$$

where

$$\mathcal{S}_t^i = \left\{ x \in \mathcal{X}_t \;\middle|\; \min_{\boldsymbol{\theta} \in \mathcal{C}_t^i} \langle \boldsymbol{\theta}, x \rangle \geq (1-\alpha)\mu^i(x_{b_t}) \right\}.$$

Here, $\mathcal{C}_t^i = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \leq c_t\}$ is the confidence region for objective i. Intuitively, $\mathcal{S}_t^i$ includes arms whose worst-case reward is guaranteed to exceed a fraction $(1-\alpha)$ of the baseline's reward for objective i. By intersecting these sets, $\mathcal{S}_t$ contains only arms that are simultaneously safe across all objectives at round t. This differs from LexUCB-C, where safety is verified cumulatively using a global reward budget, and thus the candidate arm could be risky in the short term as long as cumulative safety is maintained.

LexTS-S then checks whether the safety region $\mathcal{S}_t$ is nonempty and whether the design matrix is sufficiently well-conditioned, i.e.,

$$\lambda_{\min}(\mathrm{V}_t) \geq \left(\frac{2c_t}{\Delta_\ell + \alpha\mu_\ell} \right)^2,$$

Algorithm 4 LexTS-S

Input: $T, m, B, \alpha, \delta, \lambda, w, \Delta_\ell, \mu_\ell$

Set $\mathrm{V}_1 = \lambda \mathrm{I}$, $c_1 = \beta_1 = \sqrt{\lambda}B$, $\hat{\boldsymbol{\theta}}_1^i = \mathbf{0}$
Choose a value $\rho_1 \in (0, \frac{\mu_\ell \alpha}{B+\mu_u})$
for $t = 1$ to T **do**
  Sample $\tilde{\boldsymbol{\theta}}_t^i \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_t^i, c_t^2 \cdot \mathrm{V}_t^{-1})$
  Compute the safety region: $\mathcal{S}_t = \cap_{i=1}^m \mathcal{S}_t^i$, where

$$\mathcal{S}_t^i = \left\{ x \in \mathcal{X}_t \;\middle|\; \min_{\boldsymbol{\theta} \in \mathcal{C}_t^i} \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle \geq (1-\alpha)\mu^i(\boldsymbol{x}_{b_t}) \right\}$$

  if $\mathcal{S}_t \neq \emptyset$ and $\lambda_{\min}(\mathrm{V}_t) \geq \left(\frac{2c_t}{\Delta_\ell + \alpha\mu_\ell} \right)^2$ **then**
    $\boldsymbol{x}_t \leftarrow \textsc{SelectArm}(\tilde{\boldsymbol{\theta}}_t^i, c_t + \beta_t, \mathrm{V}_t, \mathcal{S}_t)$ *(Alg. 2)*
  else
    Sample ζ_t uniformly from the unit sphere in $\mathbb{R}^d$
    $\boldsymbol{x}_t \leftarrow (1-\rho_1)\boldsymbol{x}_{b_t} + \rho_1 \zeta_t$ *(Conservative play)*
  end if
  Play $\boldsymbol{x}_t$ and observe reward vector $[y_t^1, y_t^2, \ldots, y_t^m]$
  Update $\mathrm{V}_{t+1} = \mathrm{V}_t + \boldsymbol{x}_t \boldsymbol{x}_t^\top$
  Compute the estimators $\{\hat{\boldsymbol{\theta}}_{t+1}^i\}_{i=1}^m$ by Eq. (3)
  Compute c_{t+1} and β_{t+1} by Eq. (5)
end for

where Δ_ℓ and μ_ℓ are problem-dependent constants defined in Assumption 3. If these conditions are satisfied, LexTS-S calls the SELECTARM subroutine (Algorithm 2) to select an arm $\boldsymbol{x}_t$, which passes the sampled parameters $\{\tilde{\boldsymbol{\theta}}_t^i\}_{i=1}^m$, the confidence parameter $c_t + \beta_t$, and the feasible set $\mathcal{S}_t$. This subroutine follows the same iterative filtering mechanism as in LexUCB-C, using LEXFILTER to preserve lexicographic optimality among the arms in $\mathcal{S}_t$.

If either condition fails, either because the safety region is empty or because the confidence in parameter estimates is too low, LexTS-S takes a conservative action. Instead of playing a fixed baseline arm, it mixes the conservative baseline $\boldsymbol{x}_{b_t}$ with a randomly sampled unit vector ζ_t,

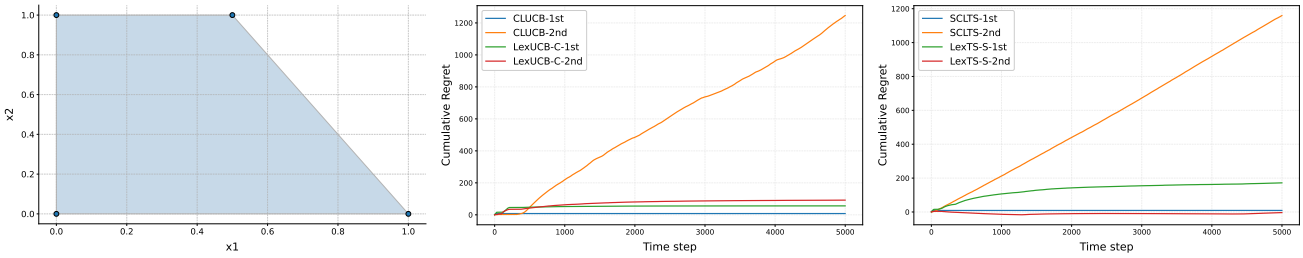
$$\boldsymbol{x}_t = (1-\rho_1)\boldsymbol{x}_{b_t} + \rho_1 \zeta_t,$$

where ρ_1 balances between baseline performance and exploratory progress. This conservative play guarantees the stage-wise safety constraint even in highly uncertain situations, similar to the conservative mechanism in LexUCB-C but with a stage-wise guarantee.

After the arm $\boldsymbol{x}_t$ is played, LexTS-S observes the reward vector $[y_t^1, y_t^2, \ldots, y_t^m]$ and updates the covariance matrix and estimators in the same manner as LexUCB-C. Finally, the confidence radius is computed as follows,

$$\begin{aligned} c_{t+1} &= \sigma \sqrt{d \log\left(\frac{mT(1+t/\lambda)}{\delta} \right)} + \sqrt{\lambda}B \\ \beta_{t+1} &= c_{t+1} \cdot \sqrt{2d \log\left(\frac{8dmT}{\delta} \right)}. \end{aligned} \tag{5}$$

In summary, LexTS-S differs from LexUCB-C in two key aspects: **(i)** it enforces stage-wise safety, requiring that each



(a) Arm Space (Convex Hull) (b) Cumulative Constraint Methods (c) Stage-wise Constraint Methods

Figure 1: Visualization of arm space and regret curves under safety constraints. (a) The arm space is a convex hull of four vertices: $\{(0,0),(1,0),(0,1),(0.5,1)\}$, where the optimal arm is $(0.5,1)$. (b) Performance comparison of CLUB and LexUCB-C under cumulative safety constraints, where LexUCB-C achieves sublinear regret on both the 1st and 2nd objectives. (c) Performance comparison of SCLTS and LexTS-S under stage-wise safety constraints, where LexTS-S achieves sublinear regret on both the 1st and 2nd objectives.

selected arm be immediately safe, rather than maintaining safety cumulatively; and **(ii)** it adopts a Thompson Sampling strategy based on posterior sampling of reward parameters, in contrast to the optimistic upper confidence bound approach used in LexUCB-C. These distinctions make LexTS-S more conservative in the short term, and better suited for environments where immediate safety is essential.

The regret bound of LexTS-S is provided as follows.

Theorem 2. *Suppose Assumptions 1 - 4 hold, and that the arm set $\mathcal{X}_t$ is convex and compact for all $t \in [T]$. Then, with probability at least $1-\delta$, for any objective $i \in [m]$, the regret of LexTS-S is bounded as*

$$R^i(T) \leq 4W^i(w) \cdot (c_T + \beta_T) \cdot \left(\sqrt{T} + 10 \log T \cdot \sqrt{dT}\right) + n_T \cdot (\Delta_u + \rho_1(\mu_u + B)),$$

where

$$W^i(w) = 1 + w + \cdots + w^{i-1},$$

$$\rho_1 = \left(\frac{\mu_\ell}{B + \mu_u}\right)\alpha, \quad h_1 = 2\rho_1(1-\rho_1) + 2\rho_1^2,$$

$$n_T \leq \left(\frac{2c_T d}{\rho_1(\Delta_\ell + \alpha\mu_\ell)}\right)^2 + \frac{2h_1^2 d^4}{\rho_1^4} \log\left(\frac{md}{\delta}\right) + \frac{2h_1 c_T d^3 \sqrt{8\log(md/\delta)}}{\rho_1^3(\Delta_\ell + \alpha\mu_\ell)}.$$

Remark 2. Theorem 2 provides a high-probability regret bound for LexTS-S under stage-wise safety constraints. The leading term $\widetilde{O}(W^i(w) \cdot d^{3/2}\sqrt{T})$ matches the typical regret rate of Thompson Sampling-based linear bandits [Abeille and Lazaric, 2017] up to a multiplicative factor $W^i(w)$, which arises from sequentially optimizing multiple objectives with lexicographic priorities. Similar to LexUCB-C, the bound ensures that the most important objective ($i = 1$) achieves nearly optimal regret scaling, i.e., $W^1(w) = 1$, while lower-priority objectives may incur increased regret due to the hierarchical optimization. The additional $d^3 \cdot \alpha^{-1}$ term reflects the cost of maintaining strict stage-wise safety throughout the learning process. In particular, a smaller α leads to more cautious behavior, ensuring that safety constraints are less likely to be violated, but at the expense of a larger regret penalty.

4 Experiments

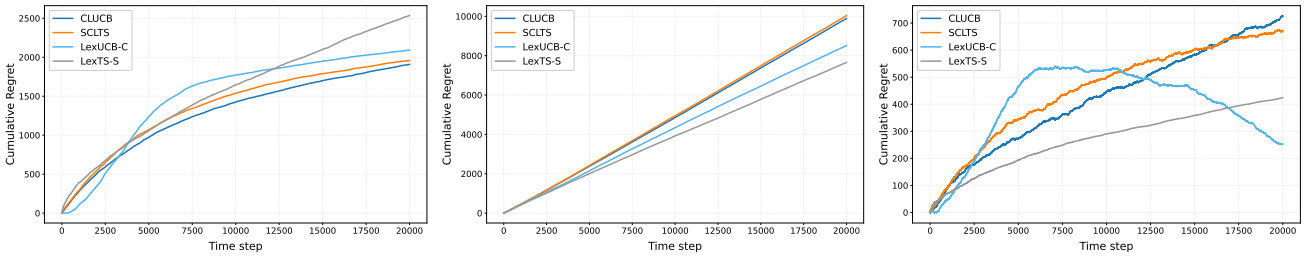
In this section, we evaluate our algorithms under safety constraints on both synthetic and real-world datasets.We summarize the baselines and briefly describe the datasets here. Full experimental details are provided in Appendix B.

We compare our methods with two representative single-objective conservative linear bandit methods: **(i) CLUCB** [Kazerouni *et al.*, 2017], which optimizes only the first objective while ensuring the *cumulative reward* stays above $(1-\alpha)$ times that of the baseline, and **(ii) SCLTS** [Moradipari *et al.*, 2020], which optimizes only the first objective and maintains the *stage-wise reward* above $(1-\alpha)$ times the baseline.

Synthetic Dataset. The decision set (arm space) is the convex hull of four vertices: $\{(0,0),(1,0),(0,1),(0.5,1)\}$, as illustrated in Figure 1(a). The inherent vectors for the two objectives are $\boldsymbol{\theta}_*^1 = [0,1]$ and $\boldsymbol{\theta}_*^2 = [1,0]$, respectively. Under this setting, the expected reward for any arm $\boldsymbol{x}$ is given by $[\langle \boldsymbol{\theta}_*^1, \boldsymbol{x}\rangle, \langle \boldsymbol{\theta}_*^2, \boldsymbol{x}\rangle] = [x_2, x_1]$. Thus, the first objective is maximized when $x_2 = 1$, achieving the optimal value of 1. All arms lying on the top edge of the convex hull satisfy this condition. Among these arms, the second objective is maximized when $x_1 = 0.5$, which yields the highest possible second-objective reward of 0.5 under the constraint that the first objective remains optimal. Therefore, the lexicographic optimal arm is $\boldsymbol{x}_t^* = [0.5, 1]$, corresponding to the expected reward vector $[1, 0.5]$. The baseline (safe) arm is set to the centroid of the convex hull, $\bar{\boldsymbol{x}} = [0.375, 0.5]$.

Under the cumulative safety constraint, Figure 1(b) shows that our proposed algorithm LexUCB-C achieves performance comparable to the baseline CLUCB on the first objective. This empirically supports our theoretical guarantee that LexUCB-C preserves the performance of the most important objective. Moreover, LexUCB-C-2nd consistently yields lower cumulative regret than CLUCB-2nd across the entire time horizon, demonstrating its effectiveness in optimizing the secondary objective without compromising safety. These results highlight the strength of LexUCB-C in balancing lexicographic reward maximization with long-term cumulative safety guarantees in multi-objective decision-making.

Under the stage-wise constraint, Figure 1(c) shows that LexTS-S incurs slightly higher regret on the primary objec-



(a) Objective 1: User Satisfaction (b) Objective 2: Diversity Promotion (c) Objective 3: Genre Alignment

Figure 2: Cumulative regret comparison between LexUCB-C and CLUCB (cumulative safety constraint) and LexTS-S and SCLTS (stage-wise safety constraint) on MovieLens 100K dataset.

tive compared to the baseline SCLTS. This is expected, as SCLTS is designed to solely optimize the primary objective. Nonetheless, the flattened regret curves of LexTS-S on both objectives suggest that it successfully identifies the lexicographic optimal arm. In contrast, SCLTS exhibits nearly linear regret on the second objective, highlighting LexTS-S's ability to handle multiple objectives simultaneously. These results demonstrate that LexTS-S effectively leverages the multi-objective structure to maintain strong performance even under stringent per-round safety constraints.

Real-world Dataset. We evaluate our proposed algorithms on the MovieLens 100K dataset[2] [Harper and Konstan, 2015], a widely used benchmark in recommender systems and bandit research [Li *et al.*, 2025]. The dataset consists of 100,000 ratings provided by 943 users on 1,682 movies, where each rating is given on a 5-star scale. Each movie is accompanied by metadata such as title, release date, and genre information covering 19 categories (Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western). We represent each movie as a 21-dimensional feature vector, consisting of a 19-dimensional one-hot encoding of its genre affiliations, a 1-dimensional normalized popularity score (ranging from 0 to 1), and a 1-dimensional bias term (set to a constant value of 1). To construct a multi-objective recommendation environment, we define three distinct objectives: User Satisfaction, Diversity Promotion, Genre Alignment.

Figures 2(a)-(c) present the cumulative regrets of all algorithms on the real-world dataset. For the first objective (Figure 2(a)), our proposed algorithms, LexUCB-C and LexTS-S, exhibit regret curves that are almost indistinguishable from those of the baselines CLUCB and SCLTS throughout the learning horizon. This observation indicates that incorporating lexicographic preferences does not compromise the optimization of the most critical objective. In the early stage, the regret of LexUCB-C slightly fluctuates due to its confidence-based exploration, while LexTS-S demonstrates smoother convergence, benefiting from the posterior sampling mechanism. As the number of rounds increases, both algorithms rapidly stabilize, achieving a comparable cumulative regret level to that of the baselines, thereby validating their ability to preserve first-priority optimality.

In contrast, for the second and third objectives (Figures 2(b)-(c)), the advantage of our lexicographic design becomes evident. Both LexUCB-C and LexTS-S achieve smaller cumulative regrets compared to CLUCB and SCLTS, with the performance gap widening over time. This demonstrates that, after optimizing the dominant objective, our algorithms effectively exploit the remaining exploration capacity to improve outcomes for subsequent objectives. The hierarchical update rules in LexUCB-C and LexTS-S enable the agent to adaptively focus on secondary and tertiary objectives without violating the lexicographic priority constraints. Overall, these results confirm that the proposed algorithms not only maintain competitiveness on the leading objective but also substantially enhance performance on subordinate ones, thereby realizing true lexicographic optimization in practice.

5 Conclusion and Future Work

In this paper, we formalize and investigate the problem of safe lexicographic MOSLB, which integrates lexicographic preferences with two widely adopted safety constraints: cumulative and stage-wise constraints. We design two algorithms tailored to these settings. For the cumulative-safe setting, we develop LexUCB-C, a UCB-based algorithm that respects lexicographic order while ensuring cumulative safety. For the stage-wise-safe setting, we offer LexTS-S, a Thompson Sampling-based method that guarantees per-round safety. We provide rigorous regret analysis for both algorithms, showing that their performance matches that of existing single-objective safe bandit algorithms [Kazerouni *et al.*, 2017; Moradipari *et al.*, 2020] in terms of dependence on the dimension d, time horizon T, and safety threshold α. Notably, we demonstrate that optimizing lower-priority objectives incurs minimal additional regret to the highest-priority objective.

This work opens several promising directions for future research. First, extending our framework to nonlinear reward models or more general action spaces would broaden its applicability. Second, while our analysis provides worst-case guarantees, exploring problem-dependent algorithms that leverage structural properties to improve empirical performance is an important direction. Finally, eliminating the reliance on prior knowledge assumed in Assumptions 1-4 would enhance the practical utility of our algorithms.

[2]https://movielens.org/

Acknowledgments

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region [GRF Project No. CityU 9043546] and the National Natural Science Foundation of China (NSFC) under Grant No. 62276223.

References

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

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 176–184, 2017.

Jacob A. Abernethy, Robert Schapire, and Umar Syed. Lexicographic optimization: Algorithms and stability. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 2503–2511, 2024.

Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems 29*, page 3458–3467, 2016.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013.

J. Aldaz, Sorina Barza, Masatoshi Fujii, and M. Moslehian. Advances in operator cauchy–schwarz inequalities and their reverses. *Annals of Functional Analysis*, 6(3):275–295, 2015.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems 32*, pages 9256 – 9266, 2019.

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

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

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

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation*, 3(1):1 – 26, 2015.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, page 207–216, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

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

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366, 2008.

Nirjhar Das and Gaurav Sinha. Linear contextual bandits with hybrid payoff: Revisited. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference*, page 441–455, 2024.

Emily Diana, Wesley Gill, Ira Globus-Harris, Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Lexicographically fair learning: Algorithms and generalization. In *2nd Symposium on Foundations of Responsible Computing*, pages 6:1–6:23, 2021.

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

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

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3962–3969, 2020.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–9, 2015.

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

Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *Proceedings of the 2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pages 202–219, 2019.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to Bayesian optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22643–22671, 2024.

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

Sumeet Katariya, Branislav Kveton, Zheng Wen, and Vamsi K. Potluru. Conservative exploration using interleaving. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 954–963, 2019.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems 30*, page 3913–3922, 2017.

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

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

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

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

Zhuohua Li, Maoli Liu, Xiangxiang Dai, and John C.S. Lui. Towards efficient conversational recommendations: Expected value of information meets bandit learning. In *Proceedings of the ACM on Web Conference 2025*, pages 4226–4238, 2025.

Jiabin Lin, Xian Yeow Lee, Talukder Jubery, Shana Moothedath, Soumik Sarkar, and Baskar Ganapathysubramanian. Stochastic conservative contextual linear bandits. In *IEEE 61st Conference on Decision and Control*, pages 7321–7326, 2022.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. In *Advances in Neural Information Processing Systems 35*, pages 16522–16532, 2022.

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

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

Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. In *Advances in Neural Information Processing Systems 33*, pages 11191–11201, 2020.

Mario Rodriguez, Christian Posse, and Ethan Zhang. Multiple objective optimization in recommender systems. In *Proceedings of the 6th ACM Conference on Recommender Systems*, pages 11–18, 2012.

Diederik M. Roijers, Luisa M. Zintgraf, and Ann Nowe. Interactive thompson sampling for multi-objective multi-armed bandits. In *Algorithmic Decision Theory: 5th International Conference*, pages 18–34, 2017.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1167–1178, 2013.

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20253–20277, 2022.

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

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

Cem Tekin. The biobjective multiarmed bandit: learning approximate lexicographic optimal allocations. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(2):1065–1080, 2019.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Nirandika Wanigasekara, Yuxuan Liang, Siong Thye Goh, Ye Liu, Joseph Jay Williams, and David S. Rosenblum. Learning multi-objective rewards and user utility function in contextual bandits for personalized ranking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3835–3841, 2019.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, page 1254–1262, 2016.

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

Ruitu Xu, Yifei Min, and Tianhao Wang. Noise-adaptive thompson sampling for linear contextual bandits. In *Advances in Neural Information Processing Systems 36*, pages 23630–23657, 2023.

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

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

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

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

Hengrui Zhang, Youfang Lin, Sheng Han, and Kai Lv. Lexicographic actor-critic deep reinforcement learning for urban autonomous driving. *IEEE Transactions on Vehicular Technology*, 72(4):4308–4319, 2023.

A Related Work

In this section, we review the related literature across three closely related research areas that are most relevant to our work: stochastic linear bandits, safety-constrained bandits, and multi-objective bandits.

A.1 Stochastic Linear Bandits

Dani *et al.* [2008] formalized the stochastic linear bandit problem and employed the confidence ellipsoid technique to derive a regret bound of $\widetilde{O}(d\sqrt{T})$. Subsequent studies refined this technique: Abbasi-yadkori *et al.* [2011] proposed the OFUL algorithm with sharper confidence bounds, and Rusmevichientong and Tsitsiklis [2010] studied improved exploration strategies under additional assumptions. Parallel to this, posterior sampling methods demonstrated superior empirical performance and achieved a regret bound of $\widetilde{O}(d^{3/2}\sqrt{T})$ [Agrawal and Goyal, 2013; Abeille and Lazaric, 2017]. Recent work has explored variations of this model, including contextual constraints [Lin *et al.*, 2022], hybrid reward structures [Das and Sinha, 2024], and lower bounds that confirm the minimax optimality of existing algorithms [Lattimore and Szepesvári, 2020]. However, these methods focus exclusively on single-objective learning and do not account for multi-objective trade-offs or safety constraints.

A.2 Safety-Constrained Bandits

Bandit optimization with safety constraints has been extensively studied [Singla and Krause, 2013; Babaioff *et al.*, 2015], with the Bandits with Knapsacks (BwK) framework [Badanidiyuru *et al.*, 2013] serving as a foundational model. This framework addresses reward maximization subject to cumulative budget or resource constraints. Early algorithms in this domain rely on primal-dual methods and linear programming [Badanidiyuru *et al.*, 2013; Agrawal and Devanur, 2016], while later work extends these techniques to adversarial settings [Immorlica *et al.*, 2019; Sivakumar *et al.*, 2022] and dynamic environments [Liu *et al.*, 2022]. A related but distinct line of constrained bandit work focuses on conservative bandits, where the learner must ensure performance remains above a fraction of a baseline policy. This idea was first introduced by Wu *et al.* [2016] in the stochastic MAB setting and later extended to linear bandits by Kazerouni *et al.* [2017] and Garcelon *et al.* [2020], with methods typically relying on modified confidence sets or feasibility projections to maintain cumulative safety. Another important direction involves stage-wise constraints, which require safety to be satisfied at every round, reflecting stricter guarantees suitable for sensitive applications such as healthcare or autonomous systems. In this setting, Amani *et al.* [2019] and Moradipari *et al.* [2020] developed UCB- and TS-based algorithms for linear bandits. However, these works are limited to single-objective settings and do not address the more complex challenges of balancing multiple objectives or respecting strict preference hierarchies.

A.3 Multi-Objective Bandits

The study of multi-objective bandits began with the work of Drugan and Nowe [2013], who proposed to handle vector-valued rewards by scalarizing them into a single-objective form. Subsequent works explored more flexible scalarization strategies, including adaptive and non-linear scalarization methods [Roijers *et al.*, 2017; Wanigasekara *et al.*, 2019]. Parallel to scalarization-based methods, another line of research focused on Pareto-optimality, either aiming to approximate the Pareto front [Auer *et al.*, 2016; Kone *et al.*, 2024] or to minimize cumulative Pareto regret [Lu *et al.*, 2019; Xu and Klabjan, 2023]. While these approaches offer principled ways to deal with multiple objectives, they typically rely on fixed or known preference weights and cannot capture strict priority structures across objectives.

To address this limitation, lexicographic preferences have recently attracted growing attention. Tekin and Turgay [2018] first considered lexicographic ordering in contextual bandits with two objectives. This was later extended by Hüyük and Tekin [2021] to the multi-objective MAB setting, accommodating more than two objectives and analyzing priority-based regret bounds. Building on this, Xue *et al.* [2024] investigated the lexicographic Lipschitz bandit problem. Beyond bandits, lexicographic optimization has also been explored beyond the bandit domain, including in multi-objective optimization [Abernethy *et al.*, 2024], reinforcement learning [Skalse *et al.*, 2022; Zhang *et al.*, 2023], and fairness-aware learning [Diana *et al.*, 2021]. Despite these advances, lexicographic bandits under safety constraints remain largely unexplored.

B Experimental Settings

This section describes the experimental settings used to evaluate the proposed algorithms, including the computational environment, synthetic data generation, and real-world recommendation scenarios. All experiments are conducted on a Windows 10 machine equipped with an Intel(R) Core(TM) i7-11700 CPU.

B.1 Synthetic Experimental Setting

We set the time horizon to $T = 5000$, the regularization parameter to $\lambda = 1.0$, the safety tolerance to $\alpha = 0.2$, and the confidence level to $\delta = 0.05$. At each round $t \in [T]$, the stochastic reward for objective i is sampled from a Gaussian distribution with mean $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t \rangle$ and variance 0.1. To ensure statistical robustness, each experimental setting is repeated 10 times with different random seeds, and we report the averaged results. Following prior work [Chapelle and Li, 2011; Jun *et al.*, 2017], we scale the confidence terms in all algorithms by a tuning factor selected from the range $[0.01, 1]$.

B.2 Real-world Experimental Setting

In the real-world experiments, we select the 500 movies (arms) with the highest number of user-assigned ratings and a subset of 20 users who have provided the most ratings. To mitigate user-specific biases, we maintain separate estimators for each user, ensuring that the learning process for one user does not affect the performance estimates of others.

To construct a multi-objective recommendation environment, we define three distinct objectives that represent different aspects of recommendation quality:

- Objective 1 (User Satisfaction): The normalized rating score, computed as $r/5.0$, where r denotes the original 1-5 star rating. This serves as the primary optimization target for conventional single-objective algorithms.
- Objective 2 (Diversity Promotion): A popularity-based term, defined as $1 - \frac{\text{popularity}(\boldsymbol{x})}{\max(\text{popularity})}$, where popularity$(\boldsymbol{x})$ is the number of ratings received by movie $\boldsymbol{x}$. This objective promotes exposure to less popular items, mitigating the filter bubble effect.
- Objective 3 (Genre Alignment): The cosine similarity between the user's preference vector and the movie's genre vector. User preferences are derived by averaging the genre vectors of movies previously rated by the user, reflecting their genre affinity. This objective ensures that recommendations align with individual content preferences.

All algorithms are configured with the following hyperparameters: $T = 20{,}000$, $\alpha = 0.3$, $\lambda = 1.0$, and $\delta = 0.05$. Experiments are conducted using different random seeds, and results are averaged over 10 runs to ensure statistical reliability. We employ a comprehensive evaluation framework assessing performance across all three objectives. For each user, we define a baseline movie as the one with their highest historical average rating (Objective 1). At each time step t, the experimental procedure proceeds as follows:

1. A user is sampled randomly from the pool of 20 users.
2. Each algorithm independently selects a movie according to its specific policy, given the current user context.
3. Upon arm selection, the true multi-objective reward vector is retrieved from the historical rating data, representing: user satisfaction, diversity measure, and genre alignment for the chosen movie-user pair.
4. Each algorithm updates its internal parameters based on the observed rewards.
5. Per-objective regrets are then computed by comparing the obtained rewards against those of the user-specific optimal movies, i.e., the movies that would have achieved the lexicographic optimal reward.

This experimental protocol ensures a consistent evaluation across users and objectives, and facilitates a clear comparison between multi-objective algorithms and single-objective conservative baselines under both performance and safety considerations.

C Proof of Theorem 1

Recall that $\mathcal{T}_T \subseteq [T]$ denote the set of rounds in which the algorithm plays an arm from the optimistic set $\mathcal{X}_t$, and let $\mathcal{T}_T^c = [T] \setminus \mathcal{T}_T$ be the rounds where the algorithm falls back to the baseline action $\boldsymbol{x}_{b_t}$ due to the safety constraint. Then, for any objective $i \in [m]$, we decompose the regret as:

$$\begin{aligned} R^i(T) &= \sum_{t \in \mathcal{T}_{T+1}} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_t, \boldsymbol{\theta}_*^i \rangle + \sum_{t \in \mathcal{T}_{T+1}^c} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_{b_t}, \boldsymbol{\theta}_*^i \rangle \\ &\leq \underbrace{\sum_{t \in \mathcal{T}_T} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_t, \boldsymbol{\theta}_*^i \rangle}_{R^i(\mathcal{T}_T)} + \underbrace{n_T \cdot \Delta_u}_{R^i(\mathcal{T}_T^c)}, \end{aligned}$$

where $n_T := |\mathcal{T}_T^c|$ is the number of fallback rounds, and Δ_u denotes the maximum suboptimality gap on any objective as we given in Assumption 3.

Bounding the Optimistic Rounds $R^i(\mathcal{T}_T)$. We bound the regret incurred in rounds $t \in \mathcal{T}_T$, which highly depends on the structure of lexicographic elimination.

Lemma 1. *With probability at least $1 - \delta$, for any $\boldsymbol{x} \in \mathcal{X}_t$, $|\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle| \leq c_t \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}, \forall i \in [m], \forall t \geq 1$.*

Proof. The σ-sub-Gaussian property of the rewards allows us to extend the Confidence Ellipsoid theorem of Abbasi-yadkori *et al.* [2011] to the multi-objective context. Precisely, the Confidence Ellipsoid theorem of Abbasi-yadkori *et al.* [2011] states that for a fixed $i \in [m]$, $\boldsymbol{\theta}_*^i$ lies in the confidence region

$$\mathcal{C}_t^i = \left\{ \boldsymbol{\theta} \;\middle|\; \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \leq \sigma \sqrt{d \log\left(\frac{1 + |\mathcal{S}_t|/\lambda}{\delta} \right)} + \sqrt{\lambda} B \right\}$$

with probability at least $1-\delta$. Then, taking a union bound over all m objectives, we get that with probability at least $1-\delta$, for any $i \in [m]$, $\boldsymbol{\theta}_*^i$ lies in the confidence region

$$\mathcal{C}_t^i = \left\{ \boldsymbol{\theta} \;\middle|\; \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \le \sigma \sqrt{d \log\left(\frac{m(1+|\mathcal{S}_t|/\lambda)}{\delta}\right)} + \sqrt{\lambda} B := c_t \right\}.$$

Through the Lagrange method [Boyd and Vandenberghe, 2004], we calculate the upper confidence bound of $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle$ for a given arm $\boldsymbol{x} \in \mathcal{X}_t$ as

$$u_t^i(\boldsymbol{x}) = \max_{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \le c_t} \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle = \max_{\|\boldsymbol{\theta}\|_{\mathrm{V}_t} \le c_t} \langle \boldsymbol{\theta} + \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle = \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + c_t \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

Similarly, the lower confidence bound of $\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle$ for an arm $\boldsymbol{x} \in \mathcal{X}_t$ is given by

$$\ell_t^i(\boldsymbol{x}) = \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - c_t \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

Thus, we can conclude that with probability at least $1-\delta$, for any $\boldsymbol{x} \in \mathcal{X}_t$, the following inequality holds for all $i \in [m]$ and $t \ge 1$,

$$|\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle| \le c_t \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

The proof of Lemma 1 is finished. □

Next, we provide a high-probability bound on how far the arms in the lexicographic filtering set $\mathcal{X}_{t,s}^m$ deviate from the true lexicographic optimal arm $\boldsymbol{x}_t^*$. This is formalized in the following lemma.

Lemma 2. *In Algorithm 3, if $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}$ and $\|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} \le C$ for any $\boldsymbol{x} \in \mathcal{X}_{t,s}$, then with probability at least $1-\delta$, $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^m$ and for any $\boldsymbol{x} \in \mathcal{X}_{t,s}^m$,*

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

Proof. We first recall the confidence bounds. Since $\mathcal{X} \subseteq \mathcal{X}_{t,s}$, then with probability at least $1-\delta$, for any $\boldsymbol{x} \in \mathcal{X}$ and $i \in [m]$, it holds that

$$\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + c_t \cdot C \ge \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle \ge \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - c_t \cdot C.$$

We proceed by induction on the objective index $i \in [m]$.

Base case ($i = 1$): By assumption, $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^0 = \mathcal{X}_{t,s}$. The first-stage filtered set is defined as

$$\mathcal{X}_{t,s}^1 := \left\{ \boldsymbol{x} \in \mathcal{X}_{t,s}^0 \;\middle|\; \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + c_t \cdot C \ge \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^1 \rangle - c_t \cdot C \right\}.$$

Because $\boldsymbol{x}_t^*$ is the lex-optimal arm and both $\boldsymbol{x}_t^*$ and $\hat{\boldsymbol{x}}_t^1$ are in $\mathcal{X}_{t,s}$, we have

$$\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle + c_t \cdot C \ge \langle \boldsymbol{\theta}_*^1, \boldsymbol{x}_t^* \rangle \ge \langle \boldsymbol{\theta}_*^1, \hat{\boldsymbol{x}}_t^1 \rangle \ge \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^1 \rangle - c_t \cdot C,$$

so $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^1$. For any $\boldsymbol{x} \in \mathcal{X}_{t,s}^1$, we have

$$\begin{aligned} \langle \boldsymbol{\theta}_*^1, \boldsymbol{x} \rangle &\ge \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - c_t \cdot C \ge \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^1 \rangle - 3c_t \cdot C \\ &\ge \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle - 3c_t \cdot C \ge \langle \boldsymbol{\theta}_*^1, \boldsymbol{x}_t^* \rangle - 4c_t \cdot C. \end{aligned}$$

Thus,

$$\langle \boldsymbol{\theta}_*^1, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle \le 4c_t \cdot C.$$

Inductive step: Suppose for $i - 1 \ge 1$ we have

$$\begin{aligned} &\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^{i-1}, \quad \text{and} \quad \forall \boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1},\ j \in [i-1], \\ &\langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle \le 4(1 + w + \cdots + w^{j-1}) \cdot c_t \cdot C. \end{aligned}$$

Let $\hat{\boldsymbol{x}}_t^i := \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1}} \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle$. Then from the lex-smoothness assumption, we get

$$\begin{aligned} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* \rangle &\ge \langle \boldsymbol{\theta}_*^i, \hat{\boldsymbol{x}}_t^i \rangle - w \cdot \max_{j \in [i-1]} \langle \boldsymbol{\theta}_*^j, \boldsymbol{x}_t^* - \hat{\boldsymbol{x}}_t^i \rangle \\ &\ge \langle \boldsymbol{\theta}_*^i, \hat{\boldsymbol{x}}_t^i \rangle - 4w \cdot (1 + w + \cdots + w^{i-2}) \cdot c_t \cdot C. \end{aligned}$$

Since

$$\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle + c_t \cdot C \ge \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* \rangle, \quad \langle \boldsymbol{\theta}_*^i, \hat{\boldsymbol{x}}_t^i \rangle \ge \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i \rangle - c_t \cdot C,$$

we conclude

$$\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle \geq \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i \rangle - 2c_t \cdot C - 4w(1 + w + \cdots + w^{i-2}) \cdot c_t \cdot C.$$

This gives

$$\langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle \geq \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i \rangle - (2 + 4w + 4w^2 + \cdots + 4w^{i-1}) \cdot c_t \cdot C,$$

and thus $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^i$ with

$$\mathcal{X}_{t,s}^i = \left\{ \boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1} \middle| \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle \geq \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i \rangle - (2 + 4w + \cdots + 4w^{i-1}) \cdot c_t \cdot C \right\}.$$

Finally, since $\hat{\boldsymbol{x}}_t^i := \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}_{t,s}^{i-1}} \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle$ and $\boldsymbol{x}_t^* \in \mathcal{X}_{t,s}^{i-1}$, for any $\boldsymbol{x} \in \mathcal{X}_{t,s}^i$, we have

$$\begin{aligned}
\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle &\geq \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle - c_t \cdot C \\
&\geq \langle \hat{\boldsymbol{\theta}}_t^i, \hat{\boldsymbol{x}}_t^i \rangle - c_t \cdot C - (2 + 4w + \cdots + 4w^{i-1}) \cdot c_t \cdot C \\
&\geq \langle \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x}_t^* \rangle - c_t \cdot C - (2 + 4w + \cdots + 4w^{i-1}) \cdot c_t \cdot C \\
&\geq \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* \rangle - 4(1 + w + \cdots + w^{i-1}) \cdot c_t \cdot C.
\end{aligned}$$

So

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle \leq 4(1 + w + \cdots + w^{i-1}) \cdot c_t \cdot C.$$

This completes the induction and proves the lemma. □

Lemma 3. *In Algorithm 2, with probability at least $1 - \delta$, for any $s \geq 1$ and $\boldsymbol{x} \in \mathcal{X}_{t,s}$,*

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

Proof. For $s = 1$, Lemma 3 holds obviously since $\|\boldsymbol{\theta}_*^i\| \leq B$ and $\|\boldsymbol{x}\| \leq 1$ for any $\boldsymbol{x} \in \mathcal{X}_{t,s}$. For $s \geq 2$, note that

$$\mathcal{X}_{t,s} = \textsc{LexFilter}(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s-1}, 2^{-s+1}),$$

we consider to utilize Lemma 2, which is the theoretical guarantee for LexFilter. However, Lemma 2 holds under the assumption that the optimal arm x_t^* belongs to the input arm set $\mathcal{X}_{t,s-1}$. Therefore, we first prove that $x_t^* \in \mathcal{X}_{t,s}$ for any $s \geq 1$.

We take use of the induction method with respect to the stage index s to prove it. For $s = 1$, $x_t^* \in \mathcal{X}_{t,1}$ obviously since $\mathcal{X}_{t,1} = \mathcal{X}_t$. For $s \geq 1$, if $x_t^* \in \mathcal{X}_{t,s}$, then Lemma 2 guarantees that $x_t^* \in \mathcal{X}_{t,s+1}$. Thus, $x_t^* \in \mathcal{X}_{t,s}$ for all $s \geq 1$.

With $x_t^* \in \mathcal{X}_{t,s-1}$ and $\mathcal{X}_{t,s} = \textsc{LexFilter}(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s-1}, 2^{-s+1})$, we can easily finish the proof of Lemma 3 by Lemma 2. □

Lemma 4. *In Algorithm 2, with probability at least $1 - \delta$, for any $\boldsymbol{x} \in \mathcal{X}_{t,T}$,*

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x} \rangle \leq 4(1 + w + \cdots + w^{i-1}) \cdot c_t \cdot \frac{1}{\sqrt{T}}, i \in [m].$$

Proof. The proof of Lemma 2 indicates that $x_t^* \in \mathcal{X}_{t,s}$ for any $s \geq 1$. Due to $\mathcal{X}_{t,T} = \textsc{LexFilter}(\{\hat{\boldsymbol{\theta}}_t^i\}_{i=1}^m, c_t, \mathcal{X}_{t,s}, 1/\sqrt{T})$, taking $C = 1/\sqrt{T}$ into Lemma 2 finishes this proof. □

We now bound the number of rounds in which the confidence interval width exceeds a given resolution level. This result is essential for controlling the cumulative regret contribution from uncertain decisions.

Lemma 5. *In Algorithm 2, define*

$$\psi_s(T) := \left\{ t \in [T] \,\middle|\, \|\boldsymbol{x}_t\|_{\mathrm{V}_t} > 2^{-s} \right\},$$

i.e., the set of rounds where the confidence width exceeds 2^{-s}. Then, for any $s \in [S]$, it holds that

$$|\psi_s(T)| \leq 5 \cdot 2^s \sqrt{d|\psi_s(T)| \log |\psi_s(T)|}.$$

Proof. Let $\widetilde{\mathrm{V}}_t := \mathrm{I} + \sum_{\tau \in \psi_s(t-1)} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top$ denote the design matrix restricted to the past rounds in $\psi_s(t-1)$. By Lemma 3 of Chu *et al.* [2011], we have:

$$\sum_{t \in \psi_s(T)} \|\boldsymbol{x}_t\|_{\widetilde{\mathrm{V}}_t^{-1}} \leq 5\sqrt{d|\psi_s(T)| \log |\psi_s(T)|}.$$

Since $\psi_s(t-1) \subseteq [t-1]$, it follows that $\widetilde{\mathrm{V}}_t \preceq \mathrm{V}_t$, and hence

$$\|\boldsymbol{x}_t\|_{\mathrm{V}_t^{-1}} \leq \|\boldsymbol{x}_t\|_{\widetilde{\mathrm{V}}_t^{-1}}.$$

Substituting this into the previous inequality yields:

$$\sum_{t\in\psi_s(T)} \|\boldsymbol{x}_t\|_{\mathrm{V}_t^{-1}} \le 5\sqrt{d|\psi_s(T)|\log|\psi_s(T)|}.$$

By the definition of $\psi_s(T)$, for all $t \in \psi_s(T)$, the confidence width satisfies $\|\boldsymbol{x}_t\|_{\mathrm{V}_t^{-1}} > 2^{-s}$. Using this, we obtain:

$$2^{-s}|\psi_s(T)| < \sum_{t\in\psi_s(T)} \|\boldsymbol{x}_t\|_{\mathrm{V}_t^{-1}} \le 5\sqrt{d|\psi_s(T)|\log|\psi_s(T)|}.$$

Multiplying both sides of the above inequality by 2^s completes the proof. □

Now, we are ready to bound the term $R^i(\mathcal{T}_T)$. Let $\psi_0(T) = \mathcal{T}_T \setminus \bigcup_{s\in[S]} \psi_s(T)$ denote the trials whose confidence interval width is less than or equal to $1/\sqrt{T}$. Then, for any objective $i \in [m]$, its regret can be rewritten as

$$R^i(\mathcal{T}_T) = \sum_{t\in\psi_0(T)} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t\rangle + \sum_{s=1}^{S}\sum_{t\in\psi_s(T)} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t\rangle \tag{6}$$

Lemma 4 tells that

$$\sum_{t\in\psi_0(T)} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t\rangle \le |\psi_0(T)| \cdot 4(1 + w + \cdots + w^{i-1}) \cdot c_T \cdot \frac{1}{\sqrt{T}}. \tag{7}$$

Lemma 3 tells that

$$\sum_{t\in\psi_s(T)} \langle \boldsymbol{\theta}_*^i, \boldsymbol{x}_t^* - \boldsymbol{x}_t\rangle \le |\psi_s(T)| \cdot 4(1 + w + \cdots + w^{i-1}) \cdot c_T \cdot 2^{-s+1}. \tag{8}$$

Take Eq. (7) and Eq. (8) into Eq. (6), the regret for the i-the objective can be bounded as

$$R^i(\mathcal{T}_T) \le 4(1 + w + \cdots + w^{i-1}) \cdot c_T \cdot \left(\frac{|\psi_0(T)|}{\sqrt{T}} + \sum_{s=1}^{S} 2 \cdot 2^{-s}|\psi_s(T)|\right). \tag{9}$$

By Lemma 5 and the fact $|\psi_s(T)| \le T$, we obtain

$$\sum_{s=1}^{S} 2 \cdot 2^{-s}|\psi_s(T)| \le 10\sum_{s=1}^{S}\sqrt{d|\psi_s(T)|\log(T)}.$$

Next, a simple application of the Cauchy-Schwartz inequality [Aldaz *et al.*, 2015] relaxes the above inequality as

$$\sum_{s=1}^{S} 2 \cdot 2^{-s}|\psi_s(T)| \le 10\sqrt{dST\log(T)}.$$

Due to $S = \lfloor \log(T) \rfloor$, we can further relax it as

$$\sum_{s=1}^{S} 2 \cdot 2^{-s}|\psi_s(T)| \le 10\log(T)\sqrt{dT}. \tag{10}$$

Taking Eq. (10) into Eq. (9) shows that

$$R^i(\mathcal{T}_T) \le 4(1 + w + \cdots + w^{i-1}) \cdot c_T \cdot \left(\sqrt{T} + 10\log(T)\sqrt{dT}\right).$$

Bounding the Conservative Rounds $R^i(\mathcal{T}_T^c)$. To control the conservative term $n_T \cdot \Delta_h$, we borrow the idea from the conservative bandit literature. Precisely, Theorem 5 of Kazerouni *et al.* [2017] states that if $\lambda \le 1$ and $\boldsymbol{\theta}_*^i \in \mathcal{C}_t^i$ for all $i \in [m]$ and $t \ge 1$, then

$$n_T \le 1 + 114d^2\frac{(B\sqrt{\lambda}+\sigma)^2}{\alpha\mu_\ell \cdot (\Delta_\ell + \alpha\mu_\ell)}\left[\log\left(\frac{64d\sqrt{m}(B\sqrt{\lambda}+\sigma)}{\sqrt{\delta}(\Delta_\ell + \alpha\mu_\ell)}\right)\right]^2.$$

Conclusion. Based on the upper bounds on $R^i(\mathcal{T}_T)$ and $R^i(\mathcal{T}_T^c)$, we can conclude that the regret of Algorithm 1 is

$$\begin{aligned} R^i(T) &\le 4(1 + w + \cdots + w^{i-1}) \cdot c_T \cdot \left(\sqrt{T} + 10\log(T)\sqrt{dT}\right) + \frac{d^2K\Delta_u}{\alpha\mu_\ell \cdot (\Delta_\ell + \alpha\mu_\ell)} \\ &= \widetilde{O}\left((1 + w + \cdots + w^{i-1}) \cdot d\sqrt{T} + d^2 \cdot \alpha^{-2}\right) \end{aligned}$$

where

$$K = 228(B\sqrt{\lambda}+\sigma)^2\left[\log\left(\frac{62d\sqrt{m}(B\sqrt{\lambda}+\sigma)}{\sqrt{\delta}(\Delta_\ell + \alpha\mu_\ell)}\right)\right]^2.$$

The proof of Theorem 1 is finished. □

D Proof of Theorem 2

Similar to the proof of Theorem 1, we can decompose the regret into two case. Let $\mathcal{T}_T \subseteq [T]$ denote the set of rounds in which the algorithm plays an arm returned by SELECTARM, and let $\mathcal{T}_T^c = [T] \setminus \mathcal{T}_T$ be the rounds where the algorithm falls back to the baseline action $\boldsymbol{x}_{b_t}$ due to the safety constraint. Then, for any objective $i \in [m]$, we decompose the regret as:

$$\begin{aligned} R^i(T) &= \sum_{t \in \mathcal{T}_{T+1}} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_t, \boldsymbol{\theta}_*^i \rangle + \sum_{t \in \mathcal{T}_{T+1}^c} \langle \boldsymbol{x}_t^* - ((1-\rho_1)\boldsymbol{x}_{b_t} + \rho_1 \zeta_t), \boldsymbol{\theta}_*^i \rangle \\ &= \sum_{t \in \mathcal{T}_{T+1}} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_t, \boldsymbol{\theta}_*^i \rangle + \sum_{t \in \mathcal{T}_{T+1}^c} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_{b_t} + \rho_1(\boldsymbol{x}_{b_t} - \zeta_t), \boldsymbol{\theta}_*^i \rangle \\ &\leq \underbrace{\sum_{t \in \mathcal{T}_T} \langle \boldsymbol{x}_t^* - \boldsymbol{x}_t, \boldsymbol{\theta}_*^i \rangle}_{R^i(\mathcal{T}_T)} + \underbrace{n_T \cdot (\Delta_u + \rho_1(\mu_u + B))}_{R^i(\mathcal{T}_T^c)}, \end{aligned}$$

where $n_T := |\mathcal{T}_T^c|$ is the number of conservative rounds, Δ_u is the maximum suboptimality gap, and μ_u is the maximum expected rewards of baseline arms as we given in Assumption 3.

Bounding the Optimistic Rounds $R^i(\mathcal{T}_T)$. We begin by presenting a lemma that establishes the confidence region for all estimators $i \in [m]$.

Lemma 6. *With probability at least $1-\delta$, for any $i \in [m]$ and $t \in [T]$,*

$$\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|_{\mathrm{V}_t} \leq c_t = \sigma \sqrt{d \log\left(\frac{mT(1+(t-1)/\lambda)}{\delta}\right)} + \sqrt{\lambda} B,$$

$$\|\tilde{\boldsymbol{\theta}}_t^i - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \leq \beta_t = c_t \cdot \sqrt{2d \log\left(\frac{8dmT}{\delta}\right)}.$$

Proof. For a fixed objective $i \in [m]$, Lemma 1 of Abeille and Lazaric [2017] guarantees that, with probability at least $1-\delta$, for any round $t \in [T]$,

$$\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|_{\mathrm{V}_t} \leq \sigma \sqrt{d \log\left(\frac{T(1+(t-1)/\lambda)}{\delta}\right)} + \sqrt{\lambda} B := \tilde{c}_t,$$

$$\|\tilde{\boldsymbol{\theta}}_t^i - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \leq \tilde{c}_t \cdot \sqrt{2d \log\left(\frac{8dT}{\delta}\right)}.$$

Applying a union bound over all $i \in [m]$ and replacing δ with δ/m finish the proof of Lemma 6. □

Based on this, the posterior rewards can be bounded as follows.

Lemma 7. *With probability at least $1-\delta$, for any $i \in [m]$ and $t \in [T]$,*

$$|\langle \boldsymbol{\theta}_*^i - \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle| \leq (c_t + \beta_t) \cdot \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

Proof. We first reformulate the expected reward as follows,

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle = \langle \boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + \langle \hat{\boldsymbol{\theta}}_t^i - \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + \langle \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle.$$

Applying the Cauchy-Schwarz inequality [Aldaz *et al.*, 2015] shows that

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle - \langle \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle \leq \|\boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} + \|\hat{\boldsymbol{\theta}}_t^i - \tilde{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

According to Lemma 6, the inequality can be further relaxed to:

$$\langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle - \langle \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle \leq (c_t + \beta_t) \cdot \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}}.$$

A similar discussion derives that

$$\begin{aligned} \langle \tilde{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle &= \langle \tilde{\boldsymbol{\theta}}_t^i - \hat{\boldsymbol{\theta}}_t^i, \boldsymbol{x} \rangle + \langle \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle + \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle \\ &\leq \|\tilde{\boldsymbol{\theta}}_t^i - \hat{\boldsymbol{\theta}}_t^i\|_{\mathrm{V}_t} \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} + \|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|_{\mathrm{V}_t} \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} + \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle \\ &\leq (c_t + \beta_t) \cdot \|\boldsymbol{x}\|_{\mathrm{V}_t^{-1}} + \langle \boldsymbol{\theta}_*^i, \boldsymbol{x} \rangle. \end{aligned}$$

Thus, the proof of Lemma 7 is complete. □

Using Lemmas 6 and 7, we can adapt the results of Lemmas 2, 3, 4 and 5 by replacing c_t with $c_t + \beta_t$ and construct the counterparts for Algorithm 4. Thus, as in the proof of Theorem 1, the regret over optimistic rounds is bounded by

$$R^i(\mathcal{T}_T) \leq 4(1 + w + \cdots + w^{i-1}) \cdot (c_T + \beta_T) \cdot \left(\sqrt{T} + 10\log(T)\sqrt{dT}\right).$$

Bounding the Conservative Rounds $R^i(\mathcal{T}_T^c)$. To control the conservative term, we borrow the idea from the stage-wise conservative bandit literature. Precisely, Theorem 3.3 of Moradipari *et al.* [2020] states that if $\lambda \geq 1$ and $\boldsymbol{\theta}_*^i \in \mathcal{C}_t^i$ for all $i \in [m]$ and $t \geq 1$, then

$$n_T \leq \left(\frac{2c_T d}{\rho_1(\Delta_\ell + \alpha\mu_\ell)}\right)^2 + \frac{2h_1^2 d^4}{\rho_1^4}\log\left(\frac{md}{\delta}\right) + \frac{2h_1 c_T d^3\sqrt{8\log(md/\delta)}}{\rho_1^3(\Delta_\ell + \alpha\mu_\ell)}$$

where $h_1 = 2\rho_1(1-\rho_1) + 2\rho_1^2$ and $\rho_1 = (\frac{\mu_\ell}{B+\mu_u})\alpha$.

Conclusion. Based on the upper bounds on $R^i(\mathcal{T}_T)$ and $R^i(\mathcal{T}_T^c)$, we can conclude that the regret of Algorithm 4 is

$$\begin{aligned} R^i(T) &\leq 4(1 + w + \cdots + w^{i-1}) \cdot (c_T + \beta_T) \cdot \left(\sqrt{T} + 10\log(T)\sqrt{dT}\right) \\ &\quad + n_T(\Delta_u + \rho_1(\mu_u + B)) \\ &= \widetilde{O}(W^i(w)d^{3/2}\sqrt{T} + d^3 \cdot \alpha^{-1}). \end{aligned}$$

The proof of Theorem 2 is finished. □