Machine Learning (2021) 110:1233–1266
https://doi.org/10.1007/s10994-021-05956-1

Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives

Alihan Hüyük[1] · Cem Tekin[2]

Received: 3 April 2020 / Revised: 3 April 2020 / Accepted: 8 February 2021 /
Published online: 5 May 2021

Abstract

We consider multi-objective multi-armed bandit with (i) lexicographically ordered and (ii) satisficing objectives. In the first problem, the goal is to select arms that are lexicographic optimal as much as possible without knowing the arm reward distributions beforehand. We capture this goal by defining a multi-dimensional form of regret that measures the loss due to not selecting lexicographic optimal arms, and then, propose an algorithm that achieves $\tilde{O}(T^{2/3})$ gap-free regret and prove a regret lower bound of $\Omega(T^{2/3})$. We also consider two additional settings where the learner has prior information on the expected arm rewards. In the first setting, the learner only knows for each objective the lexicographic optimal expected reward. In the second setting, it only knows for each objective a near-lexicographic optimal expected reward. For both settings, we prove that the learner achieves expected regret uniformly bounded in time. Then, we show that the algorithm we propose for the second setting of lexicographically ordered objectives with prior information also attains bounded regret for satisficing objectives. Finally, we experimentally evaluate the proposed algorithms in a variety of multi-objective learning problems.

Keywords Multi-armed bandit · Multi-objective learning · Lexicographic optimality · Satisficing

Most of the work was performed while the first author was at Bilkent University.

Editor: Alan Fern.

✉ Alihan Hüyük
ah2075@cam.ac.uk

Cem Tekin
cemtekin@ee.bilkent.edu.tr

[1] University of Cambridge, Cambridge, UK

[2] Bilkent University, Ankara, Turkey

1 Introduction

A vast number of decision-making and learning tasks involve multi-dimensional performance metrics (objectives). Examples include recommending items in a recommender system to optimize accuracy, diversity and novelty (Zhou et al., 2010; Konstan et al., 2006), learning lexicographic optimal routing flows in wireless networks (Shah-Mansouri et al., 2009), and adjusting the dose of radiation therapy for cancer patients while prioritizing target coverage over proximity of the therapy to the organs at risk (Jee et al., 2007). In most of these problems, the learner aims to choose arms that yield high rewards in all of the objectives; however, it prefers arms that yield high rewards in the low-priority objectives only if they do not compromise the rewards in the high-priority objectives. For instance, in intensity modulated radiation treatment (IMRT) for cancer patients (Jee et al., 2007), the primary objective is to deliver sufficiently high doses of radiation to target volumes. A secondary objective is to minimize dose to normal tissues without underdosing to the target volumes.

There also exists a wide range of tasks where the learner does not prioritize objectives but it rather aims to satisfice a target value for each objective. Especially in engineering, design goals are often not formulated as optimization problems but rather formulated through specifications that the final design needs to satisfy. For instance, Cully et al., (2015) uses a multi-armed bandit (MAB) framework to develop control policies for robots that seek to prevent damage to the parts of the robot. Instead of minimizing the damage, the authors successfully use a satisficing objective to keep the damage below a critical threshold, which speeds up the learning process.

Motivated by these, in this paper, we propose two new MAB problems: multiobjective MAB with (i) lexicographically ordered objectives (Lex-MAB) and (ii) satisficing objectives (Sat-MAB).

In the Lex-MAB, the learner's priority over the objectives is formally captured by lexicographic ordering. Essentially, given D objectives indexed by the set $\mathcal{D} := [D]$, objective i has a higher priority than objective j if $i < j$.[1] This priority induces a preference over the finite set of arms denoted by $\mathcal{A}$. Formally, given two arms a and a' with the corresponding real-valued expected reward vectors $\boldsymbol{\mu}_a := (\mu_a^1, \dots, \mu_a^D)$ and $\boldsymbol{\mu}_{a'} := (\mu_{a'}^1, \dots, \mu_{a'}^D)$, we say that arm a lexicographically dominates arm a' in the first $i \le D$ objectives (written as $a \succ_{\text{lex},i} a'$) if $\mu_a^j > \mu_{a'}^j$, where $j := \min\{k \le i : \mu_a^k \ne \mu_{a'}^k\}$.[2] Here, the latter expression is succinctly expressed as $\boldsymbol{\mu}_a \succ_{\text{lex},i} \boldsymbol{\mu}_{a'}$. Based on this preference, the set of lexicographic optimal arms are defined as the ones that are not lexicographically dominated by any other arm in all D objectives, which is given as $\mathcal{A}_* := \{a \in \mathcal{A} : \boldsymbol{\mu}_{a'} \not\succ_{\text{lex},D} \boldsymbol{\mu}_a, \forall a' \in \mathcal{A}\}$.

In the Lex-MAB, at each round t, the learner selects an arm $a(t) = a$ from $\mathcal{A}$, and then, receives a D-dimensional random reward vector $\boldsymbol{r}(t) := (r^1(t), \dots, r^D(t))$ that is drawn from a fixed distribution with the expectation vector $\boldsymbol{\mu}_a$. The goal of the learner is to perform as well as an oracle which perfectly knows the set of lexicographic optimal arms and selects a lexicographic optimal arm in every round. We capture the ordering of the objectives by introducing a multi-dimensional regret measure called the lexicographic regret. As this regret notion is fundamentally different from the scalar regret notion used in the classical

[1] For a possitive integer $n \in \mathbb{Z}_+$, $[n] := \{1, \dots, n\}$.

[2] If there is no such j, then $\mu_a^k = \mu_{a'}^k$ for all $k \in [i]$, which implies that $\boldsymbol{\mu}_a$ does not lexicographically dominate $\boldsymbol{\mu}_{a'}$ in the first i objectives.

stochastic MAB (Lai & Robbins, 1985), minimizing it requires exploiting the multi-dimensional nature of the rewards and ordering of the objectives both in algorithm design and technical analysis.

This is a challenging task because simple techniques such as turning the problem into a MAB with scalar rewards by using scalarization methods from multi-objective optimization (Ehrgott, 2005) will not work since the solution of the scalarized problem may not produce lexicographic optimal arms. The problem is further complicated due to the fact that without any prior knowledge on the expected arm rewards, it is impossible to identify lexicographic optimal arms with high probability. This can be observed by considering a problem instance with arms a and a', and $D = 2$, such that a and a' have the same expected reward in objective 1, and a is the only lexicographic optimal arm. Although, in this problem, the learner can identify with high probability which arm is better in the second objective, it can never be sure about the lexicographic optimality of that arm. We call this problem the *identifiability problem*.[3] Despite the identifiability problem, we show that it is possible to achieve $\tilde{O}(T^{2/3})$ gap-free regret by learning to select near-lexicographic optimal arms. We also prove that our method is near optimal by showing a regret lower bound of $\Omega(T^{2/3})$.

The challenges described above motivates us to consider the cases where the learner has prior knowledge on the expected rewards in addition to the much more challenging prior-free case. Specifically, we consider two types of prior knowledge, which generalize the prior knowledge introduced in Bubeck et al. (2013) and Vakili and Zhao (2013) to multi-dimensional rewards. In the first case, we assume that the expected rewards of a lexicographic optimal arm are known. In the second case, we assume that near-lexicographic optimal expected rewards are known. Then, we build learning algorithms that utilize the prior information to achieve uniformly-bounded-in-time lexicographic regret for both cases.

Importantly, for the first case, we show that the regret in each objective due to selecting a suboptimal arm a is inversely proportional to the maximum of the gaps of arm a over all objectives. This shows that having prior information over multiple objectives speeds up elimination of suboptimal arms. This is analogous to the combinatorial MAB (Gai et al., 2012) in the sense that observations from one objective can help ruling out suboptimal arms in other objectives. We also prove that a similar gain appers in the second case, albeit we cannot rule out an arm performing much better than a lexicographic optimal arm in one of the objectives as suboptimal.

As our second contribution, we define the Sat-MAB as the generalization of the satisfaction-in-mean-reward problem introduced in Reverdy et al. (2017) to the multi-objective setting, and show that the algorithm that we propose for the second case of the Lex-MAB with prior information also optimally solves the Sat-MAB by achieving uniformly-bounded-in-time regret. This improves on satisfaction-in-mean-reward UCL algorithm given in Reverdy et al. (2017), which is only shown to achieve logarithmic-in-time regret for the single-objective case. Finally, we numerically evaluate the performance of our algorihtms on several multi-objective learning problems.

Rest of the paper is organized as follows. Related work is given in Sect. 2. The Lex-MAB, the lexicographic regret, types of prior information and the Sat-MAB are defined in Sect. 3. Algorithms and regret bounds for the Lex-MAB and the Sat-MAB are given in

[3] It also exists in the best arm identification problem with a fixed confidence (see Chapter 33 in Lattimore & Szepesvári, 2019).

Sect. 4. Experimental results are given in Sect. 5 followed by the concluding remarks in Sect. 6.

2 Related work

Multi-objective MAB Numerous works have investigated regret minimization in multi-objective variants of the MAB problem. For instance, Drugan and Nowe (2013) defines for each suboptimal arm its distance to the Pareto front as the Pareto suboptimality gap and the regret as the sum of the Pareto suboptimality gaps of the arms chosen by the learner. It proposes a learning algorithm that achieves $O(\log T)$ gap-dependent Pareto regret. Turgay et al. (2018) considers the multi-objective contextual MAB problem with similarity information, and extends the contextual zooming algorithm in Slivkins (2014) to minimize the Pareto regret while making fair selections among the estimated Pareto optimal arms. The proposed algorithm is shown to achieve $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ Pareto regret where d_p is the Pareto zooming dimension. In addition, Tekin and Turgay (2018) considers a biobjective contextual MAB problem with lexicographically ordered objectives. Unlike that work, we study the general case with D lexicographically ordered objectives and also consider the effect of prior information on learning.

Satisficing and thresholding MAB Locatelli et al. (2016) proposes the tresholding MAB, where the goal is to, after a set number of rounds, determine the arms with means that are higher or lower than a given threshold up to a given precision. Similarly, Reverdy et al. (2017) proposes MAB with satisficing objectives, where the goal is to minimize cumulative regret with respect to a given threshold. There, arms with means that have a "satisfying" probability of being higher than the threshold do not incur any regret. In particular, Reverdy et al. (2017) proposes an algorithm that achieves logarithmic-in-time regret for the satisfaction-in-mean-reward problem. The Sat-MAB proposed in our work generalizes this problem to the multi-objective case by introducing different satisficing thresholds for each objective. Moreover, we also propose an algorithm that achieves regret uniformly-bounded-in-time, which improves upon the one in Reverdy et al. (2017).

MAB with prior information Lai and Robbins (1985) shows that in the classical stochastic MAB problem, for any uniformly good policy, the regret grows at least logarithmically over time. As opposed to this, Lai and Robbins (1984) proves for the two-armed stochastic bandit that when the learner has prior information on the maximum expected reward μ^* and the minimum nonzero suboptimality gap Δ, there exist policies that can achieve uniformly bounded regret. This idea is further investigated in Bubeck et al. (2013), which shows that bounded regret of order $1/\Delta$ is achieved for the case with finitely many arms when the learner knows μ^* and a positive lower bound on Δ. Garivier et al. (2018) studies the case where only μ^* is known and proposes an algorithm with bounded regret of order $\log(1/\Delta)(1/\Delta)$, and also proves a lower bound of order $1/\Delta$. This paper also provides a generic tool to prove both gap-dependent and gap-independent lower bounds on the regret. Bubeck and Liu (2013) considers Thompson sampling and shows that its regret is uniformly bounded when μ^* and a positive lower bound on Δ are known. On the other hand, Vakili and Zhao (2013) considers a weaker prior information model where the learner knows a near-optimal expected reward η, which can be computed using μ^* and a positive lower bound on Δ. The proposed algorithm obtains $\sum_a \Delta_a/\delta^3$ regret, where $\delta = \mu^* - \eta < \Delta$ and Δ_a is the suboptimality gap of arm a. Mersereau et al. (2009) and Lattimore and Munos (2014) consider as prior information the knowledge of parameterized expected reward functions for each arm. In these works, the only unknown is the true parameter, which can be estimated by using reward observations from all of the arms.

Different from the works mentioned above, in this paper, we consider a multi-objective MAB problem with lexicographically ordered objectives. We design algorithms that exploit the prior information in all objectives simultaneously to rule out arms that are not lexicographic optimal. Our regret bounds match the ones in Garivier et al. (2018) and improve the ones in Vakili and Zhao (2013) for the case with a single objective.

3 Problem formulation

In this section, we formally define the Lex-MAB and the Sat-MAB.

3.1 The Lex-MAB

System model We consider rounds indexed by $t \in \{1, 2, \dots\}$. In each round t, the learner first selects an arm $a(t)$ from the finite arm set $\mathcal{A} := [A]$, and then, observes a random reward for each objective $i \in \mathcal{D} := [D]$, denoted by $r^i(t)$, which is equal to $\mu^i_{a(t)} + \kappa^i(t)$, where μ^i_a denotes the expected reward of arm a in objective i and $\kappa^i(t)$ denotes the zero mean noise. The learner does not know the expected reward vector $\boldsymbol{\mu}_a := (\mu^1_a, \dots, \mu^D_a)$ for any $a \in \mathcal{A}$ beforehand, and given $a(t) = a$, the noise vector $\{\kappa^1(t), \dots, \kappa^D(t)\}$ is sampled from a fixed (unknown) multivariate distribution $\boldsymbol{\nu}_a$, independent of the other rounds. Moreover, its marginal distribution is 1-sub-Gaussian, i.e., $\forall a \in \mathcal{A}$ and $\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda\kappa^i(t)} | a(t) = a] \leq \exp(\lambda^2/2)$.[4] The assumption on the noise distribution is very general as it covers the Gaussian distribution with zero mean and unit variance, and any bounded zero mean distribution defined over an interval of length 2.

Lexicographic optimality For two D-dimensional real-valued vectors $\boldsymbol{\mu} := (\mu^1, \dots, \mu^D)$ and $\boldsymbol{\mu}' := (\mu'^1, \dots, \mu'^D)$, and $i \in [D]$, we say that $\boldsymbol{\mu}$ lexicographically dominates $\boldsymbol{\mu}'$ in the first i objectives, denoted by $\boldsymbol{\mu} \succ_{\text{lex},i} \boldsymbol{\mu}'$, if $\mu^j > \mu'^j$, where $j := \min\{k \leq i : \mu^k \neq \mu'^k\}$. Based on this, we say that arm a lexicographically dominates arm a' in the first i objectives if $\boldsymbol{\mu}_a \succ_{\text{lex},i} \boldsymbol{\mu}_{a'}$. The complement of this is denoted by $\boldsymbol{\mu}_a \not\succ_{\text{lex},i} \boldsymbol{\mu}_{a'}$.

Let $\mathcal{A}^i_* := \{a \in \mathcal{A} : \boldsymbol{\mu}_{a'} \not\succ_{\text{lex},i} \boldsymbol{\mu}_a, \forall a' \in \mathcal{A}\}$ denote the set of lexicographic optimal arms in the first i objectives and define $\mathcal{A}_* := \mathcal{A}^D_*$. Clearly, we have $\mathcal{A}^{i+1}_* \subseteq \mathcal{A}^i_*$ for $i \in [D-1]$. We use $*$ to denote an arm that is lexicographic optimal in all objectives, and μ^i_* to denote the expected reward of this arm in objective i. Moreover, we define the *gap* of arm a in objective i as $\Delta^i_a := \mu^i_* - \mu^i_a$ and the *absolute gap* of arm a in objective i as $\nabla^i_a := |\mu^i_* - \mu^i_a|$. For $i \in \{2, \dots, D\}$, we let $\mathcal{S}^i_* := \mathcal{A}^{i-1}_* - \mathcal{A}^i_*$ denote the set of arms that are lexicographic optimal in the first $i-1$ objectives but not lexicographic optimal in the first i objectives and define $\mathcal{S}^1_* := \mathcal{A} - \mathcal{A}^1_*$. Note that $a \in \mathcal{S}^i_*$ implies that $\Delta^i_a > 0$. The set of suboptimal arms in objective i is given as $\mathcal{S}^i := \{a : \Delta^i_a > 0\}$. We also define the maximum gap in objective i as $\Delta^i_{\max} := \max_{a \in \mathcal{A}} \Delta^i_a$ and the maximum absolute gap as $\nabla^{\max}_a := \max_{j \in \mathcal{D}} \nabla^j_a$. Finally, let $\Delta^i_{\min} := \min_{a \in \mathcal{S}^i_*} \Delta^i_a$ denote the minimum gap among arms in $\mathcal{S}^i_*$.[5]

[4] Noise can be dependent over the objectives.

[5] If $\mathcal{S}^i_* = \emptyset$, then $\Delta^i_{\min} = \infty$.

Why lexicographic optimality is worth studying First, it is a very well-known concept in multi-criteria decision making (Ehrgott, 2005) and utility theory (Fishburn, 1974). Applications such as intensity modulated radiation therapy for cancer patients (Jee et al., 2007) and routing with multiple sinks (Shah-Mansouri et al., 2009) involve lexicographically ordered preferences. In an online setting, it implies that objectives have different priorities for the decision-maker and the (new) user that the decision-maker serves in each round. For instance, consider choosing treatments for patients sequentially arriving over time from the set of treatments a, b and c with expected rewards [1, 0], [1, 1] and [0, 1]. Assume that the first objective is related to effectiveness and the second objective is related to side-effects. It is not acceptable that any patient receives treatment a instead of b even when there exists a mechanism for the learner that will guarantee it to achieve a cumulative expected reward that lexicographically dominates the cumulative expected reward of what we propose. If one just tries to maximize the first objective, then it may never learn to select lexicographic optimal arms. Moreover, our results in this paper also show that learning might be faster (even in the first objective) when we use rewards from the other objectives.[6]

Types of prior knowledge:

Case 1 No prior knowledge on the expected rewards.

Case 2 Lexicographic optimal expected rewards are known, i.e., the learner knows μ^i_* for all $i \in \mathcal{D}$. For this case, we assume $\mu^i_* = 0, \forall i \in \mathcal{D}$ without any loss of generality.

Case 3 Near-lexicographic optimal expected rewards are known, i.e., the learner knows η_i such that $\mu^i_* - \Delta^i_{\text{min}} < \eta_i < \mu^i_*$ for all $i \in \mathcal{D}$. For this case, we define $\delta_i := \mu^i_* - \eta_i, \forall i \in \mathcal{D}$ and assume $\eta_i = 0, \forall i \in \mathcal{D}$ without any loss of generality.

Remark 1 In Case 2, if μ^i_*s are not equal to 0, we can subtract them from the rewards to obtain normalized rewards $\tilde{r}^i(t) := r^i(t) - \mu^i_*$. Under the normalized rewards, we will have $\tilde{\mu}^i_* = 0, \forall i \in \mathcal{D}$ and the gaps that we have defined will not be affected. Similarly, in Case 3, we can subtract η_is from the rewards to obtain $\tilde{r}^i(t) := r^i(t) - \eta_i$.

Regret definitions The (pseudo) regret of the learner is measured with respect to an oracle, which knows the expected rewards of the arms and chooses a lexicographic optimal arm in each round. We define two notions of regret: priority-based and priority-free regrets in objective i are given as

$$\text{Reg}^i_{pb}(T) := \sum_{t=1}^{T} \Delta^i_{a(t)} \mathbb{I}\{a(t) \in \mathcal{S}^i_*\}$$

and

$$\text{Reg}^i_{pf}(T) := \sum_{t=1}^{T} \Delta^i_{a(t)}$$

[6] This can be inferred from Theorems 3–6 and the comparison with the single-objective versions given in Table 3.

respectively. The lexicographic priority-based and priority-free regrets are defined as the tuples $\mathbf{Reg}_{pb}(T) := (\mathrm{Reg}^1_{pb}(T), \dots, \mathrm{Reg}^D_{pb}(T))$ and $\mathbf{Reg}_{pf}(T) := (\mathrm{Reg}^1_{pf}(T), \dots, \mathrm{Reg}^D_{pf}(T))$ respectively. Subscripts will be removed from the notation when the considered regret notion is clear from the context.

For $\mathbf{Reg}_{pb}(T)$, when $a(t) \in \mathcal{S}^i_*$, regret is incurred only in objective i. No regret is incurred for $j < i$ since $\Delta^j_{a(t)} = 0$. In addition, no regret is incurred for $j > i$ when $a(t) \in \mathcal{S}^i_*$. This definition of regret is consistent with the priority that the learner assigns to each objective. Since lexicographic ordering implies that even a small improvement in the expected reward in objective i is more important than any improvement in the expected rewards of objectives $j > i$, the learner does not care about the loss it incurs in higher indexed objectives when $a(t) \in \mathcal{S}^i_*$. For $\mathbf{Reg}_{pf}(T)$, an arm a for which $\mu^i_a > \mu^i_*$ can incur negative regret in objective i, but then, it is guaranteed that positive regret is incurred in some other objective $j < i$.

We say that the regret is $O(\max\{f_1(T), \dots, f_D(T)\})$ when $\max\{0, \mathrm{Reg}^i(T)\} \in O(f_i(T))$ for $i \in \mathcal{D}$. Under both notions of regret, the (cumulative) regret of any arm selection strategy cannot lexicographically dominate the cumulative regret of always selecting a lexicographic optimal arm, which is essentially the zero vector. Therefore, the time-averaged expected rewards of any algorithm that achieves sublinear $\mathbf{Reg}_{pb}(T)$ or $\mathbf{Reg}_{pf}(T)$ will converge (as $T \to \infty$) to the lexicographic optimal expected rewards. In addition, under $\mathbf{Reg}_{pf}(T)$ the lexicographic ordering between the cumulative expected rewards and the regrets of any pair of sequences of arms $(a(1), \dots, a(T))$ and $(a'(1), \dots, a'(T))$ will be the same.

3.2 The Sat-MAB

In this section, we extend the satisfaction-in-mean-reward problem introduced in Reverdy et al. (2017) to the multi-objective setting. We keep the same system model but introduce the concept of satisficing optimality and a new notion of regret that captures this concept.

Satisficing optimality In the satisficing setting, the learner is given a target threshold η_i for each objective $i \in \mathcal{D}$. We say that an arm a is satisficing "optimal" or simply satisficing in objective i if and only if its mean reward in objective i is equal to or larger than the corresponding target threshold. Let $\mathcal{A}^i_s := \{a \in \mathcal{A} : \mu^i_a \geq \eta_i\}$ be the set of satisficing arms in objective i and $\mathcal{S}^i_s := \mathcal{A} - \mathcal{A}^i_s$ be the set of non-satisficing arms in objective i. The satisficing goal is to play arms that are satisficing in all objectives. We assume such arms exist and call them satisficing "optimal" arms. Then, we use $*$ to denote an arbitrary satisficing "optimal" arm and call it the "optimal" satisficing arm. Note that $\eta_i \leq \mu^i_*$ for all $i \in \mathcal{D}$ and define $\delta_i := \mu^i_* - \eta_i$ for all $i \in \mathcal{D}$.

Regret definition The satisficing regret in objective i is given as $\mathrm{Reg}^i_s(T) := \sum_{t=1}^{T} (\Delta^i_{a(t)} - \delta_i)\mathbb{I}\{a(t) \in \mathcal{S}^i_s\}$ and the satisficing regret is defined as the tuple $\mathbf{Reg}_s(T) := (\mathrm{Reg}^1_s(T), \dots, \mathrm{Reg}^D_s(T))$. Note that an arm a incurs regret in objective i only when it is not satisficing in that objective and the amount of regret incurred is equal to the gap between its mean reward in objective i and the corresponding target threshold, i.e., $\Delta^i_a - \delta_i = \eta_i - \mu^i_a$.

Remark 2 When $D = 1$, the Sat-MAB reduces to the exact same problem introduced in Reverdy et al. (2017) as Problems 1 and 2 (satisfaction-in-mean-reward problem).

4 Learning algorithms and regret bounds

In this section, we propose several learning algorithms for the Lex-MAB and the Sat-MAB and analyse their regrets.

4.1 Algorithms and regret bounds for the Lex-MAB

A learning algorithm for Case 1 We propose *Prior Free Lexicographic Exploration and eXploitation* (PF-LEX) given in Algorithm 1, which learns to select near lexicographic optimal arms without any prior information on the mean arm rewards. PF-LEX takes as input $\epsilon > 0$, which is proportional to the suboptimality that the learner aims to tolerate in all objectives (this will be adjusted based on the time horizon T). For each arm a, PF-LEX keeps a counter $N_a(t)$ that counts how many times arm a was selected prior to the current round and the sample mean estimate $\hat{\mu}^i_a$ of the rewards from objective i of arm a observed prior to the current round for all $i \in \mathcal{D}$. The values of these variables at the beginning of round t are denoted by $N_a(t)$ and $\hat{\mu}^i_a(t)$ respectively.

Algorithm 1 PF-LEX

```
Input: $\epsilon$, $\delta$
Counters: $N_a, \forall a \in \mathcal{A}$
Estimates: $\hat{\mu}^i_a, \forall a \in \mathcal{A}, \forall i \in \mathcal{D}$
For each round $t$:
    Compute $u^i_a = \hat{\mu}^i_a + c_a$ and $l^i_a = \hat{\mu}^i_a - c_a$, $\forall a \in \mathcal{A}$, $\forall i \in \mathcal{D}$
    Set $\hat{a}^1_* = \text{argmax}_{a \in \mathcal{A}}\, u^1_a$, compute $\hat{\mathcal{A}}^1_* = \{a \in \mathcal{A} : a\, C_1\, \hat{a}^1_*\}$
    If there exists an arm $a$ in $\hat{\mathcal{A}}^1_*$ such that $c_a > \epsilon/2$:
        Select an arm $a(t)$ in $\hat{\mathcal{A}}^1_*$ such that $c_{a(t)} > \epsilon/2$ uniformly at random
    If all arms $a$ in $\hat{\mathcal{A}}^1_*$ satisfy $c_a \leq \epsilon/2$:
        Set $\hat{a}^i_* = \text{argmax}_{a \in \hat{\mathcal{A}}^{i-1}_*}\, u^i_a$, compute $\hat{\mathcal{A}}^i_* = \{a \in \hat{\mathcal{A}}^{i-1}_* : a\, C_i\, \hat{a}^i_*\}$, $\forall i \in \{2, \ldots, D-1\}$
        Select $a(t) = \hat{a}^D_* = \text{argmax}_{a \in \hat{\mathcal{A}}^{D-1}_*}\, u^D_a$
```

Arm selection of PF-LEX in round t depends on the confidence intervals in the first $D-1$ objectives. The *upper confidence bound* (UCB) and the *lower confidence bound* (LCB) of arm a in objective i are given as $u^i_a(t) := \hat{\mu}^i_a(t) + c_a(t)$ and $l^i_a(t) := \hat{\mu}^i_a(t) - c_a(t)$ respectively. Here,

$$c_a(t) := \sqrt{\frac{1 + N_a(t)}{N^2_a(t)}\left(1 + 2\log\left(\frac{AD\sqrt{1 + N_a(t)}}{\delta}\right)\right)}$$

represents the uncertainty in arm a's reward and δ is called the *confidence term*, which is also given as input to PF-LEX. As expected, the uncertainty decreases as arm a gets selected. It is easy to see that $\mu^i_a \in [l^i_a(t), u^i_a(t)]$ with high probability for all objectives and all rounds. In each round, PF-LEX estimates the set of near-lexicographic optimal arms. For this, similar to Joseph et al. (2016), we say that arms a and a' are *linked* in objective i if $[l^i_a(t), u^i_a(t)] \cap [l^i_{a'}(t), u^i_{a'}(t)] \neq \emptyset$. When a and a' are in the same component of the transitive closure of the linked relation in objective i, we say that they are *chained* in objective i and write $a\, C_{i,t}\, a'$. Starting from $\hat{\mathcal{A}}^0_*(t) = \mathcal{A}$, PF-LEX recursively computes the estimate $\hat{\mathcal{A}}^i_*(t)$ of $\mathcal{A}^i_*$ for $i \in [D-1]$. After it computes $\hat{\mathcal{A}}^{i-1}_*(t)$, it identifies the optimistic near-lexicographic optimal arm in objective i as $\hat{a}^i_*(t) = \operatorname{argmax}_{a \in \hat{\mathcal{A}}^{i-1}_*(t)} u^i_a(t)$. Then, it sets $\hat{\mathcal{A}}^i_*(t) = \{a \in \hat{\mathcal{A}}^{i-1}_*(t) : a\, C_{i,t}\, \hat{a}^i_*(t)\}$.

Suppose we always select $\hat{a}^D_*(t)$, which happens to be in $\mathcal{S}^i_*$ for some round t. For such rounds, we show later in Lemma 3 that the regret incurred is bounded by the length of the chain formed by $\hat{\mathcal{A}}^i_*(t)$. In order to guarantee regret that is proportional to ϵ, we want the length of the chains not to be more than a constant factor of ϵ. As it is not always possible to shrink the chains by always selecting $\hat{a}^D_*(t)$, to achieve our goal, we require all arms in $\hat{\mathcal{A}}^1_*(t)$ to have narrow confidence intervals. Thus, PF-LEX selects $a(t) = a \in \hat{\mathcal{A}}^1_*(t)$ if there is an arm a with high uncertainty, i.e., $c_a(t) > \epsilon/2$. On the other hand, if $c_a(t) \le \epsilon/2$ for all $a \in \hat{\mathcal{A}}^1_*(t)$, then PF-LEX simply selects $a(t) = \hat{a}^D_*(t)$. Algorithm 1 shows a more efficient implementation of PF-LEX that does not compute $\hat{\mathcal{A}}^j_*(t)$ for $j > 1$ when $a(t)$ is selected from $\hat{\mathcal{A}}^1_*(t)$. Finally, after PF-LEX selects arm $a(t)$, it observes the random reward vector $\boldsymbol{r}(t) = (r^1(t), \dots, r^D(t))$ of arm $a(t)$, and then, updates the sample mean estimates of the rewards in objectives $i \in \mathcal{D}$ and the counter of $a(t)$. The following theorem shows that PF-LEX achieves $\tilde{O}(T^{2/3})$ regret.

Theorem 1 *When PF-LEX is run with $\delta \in (0,1)$ and $\epsilon > 0$, with probability at least $1-\delta$, for all $i \in \mathcal{D}$ and for all $T \ge 1$, we have*

$$\mathrm{Reg}^i_{pb}(T) \le 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}^i_*|T} + \left(3 + \frac{16}{\epsilon^2}\log\frac{2\sqrt{e}AD}{\epsilon\delta}\right)|\mathcal{S}^i_*|\Delta^i_{\max} + \epsilon(A-1)T$$

where $B_{T,\delta} := \sqrt{1 + 2\log(AD\sqrt{T}/\delta)}$. Given a particular time horizon T, by setting $\epsilon = T^{-1/3}$, with probability at least $1-\delta$, we have

$$\begin{aligned}&\mathrm{Reg}^i_{pb}(T)\\ &\le 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}^i_*|T} + (A-1)T^{2/3} + \left(3 + 16T^{2/3}\log\frac{2\sqrt{e}ADT^{1/3}}{\delta}\right)|\mathcal{S}^i_*|\Delta^i_{\max}.\end{aligned}$$

Moreover, taking $\delta = 1/T$, $\mathbb{E}[\mathrm{Reg}^i_{pb}(T)] = \tilde{O}(T^{2/3})$.

Remark 3 Unlike the cases with prior information that follows, an analogue of the regret bound in Theorem 1 will not hold for the priority-free regret when $\mathcal{S}^i_*$ is replaced by $\mathcal{S}^i$. Any two arms that are both lexicographic optimal in the first $i-1$ objectives are linked in these objectives with high probability. If one happens to be the selected arm, we are confident that they are both in $\hat{\mathcal{A}}^{i-1}_*$. When the selected arm is in $\mathcal{S}^i_*$, we use this fact and compare it to a lexicographic optimal arm to conclude that the gap of the selected arm in objective i is smaller than the regret that we aim to tolerate. However, we fail to make any deductions about the higher-indexed objectives.

Proof of Theorem 1 First, we state a concentration inequality that will be used in the proof.

Lemma 1 (Lemma 6 in Abbasi-Yadkori et al., 2011) *Consider an arm a for which the rewards of objective i are generated by a process $\{R^i_a(t)\}_{t=1}^T$ with $\mu^i_a = \mathbb{E}[R^i_a(t)]$, where the noise $R^i_a(t) - \mu^i_a$ is conditionally 1-sub-Gaussian. Let $N_a(T)$ denote the number of times a is selected by the beginning of round T. Let $\hat{\mu}_a(T) = \sum_{t=1}^{T-1} \mathbb{I}\{a(t) = a\} R^i_a(t)/N_a(T)$ for $N_a(T) > 0$ and $\hat{\mu}_a(T) = 0$ for $N_a(T) = 0$. Then, for any $0 < \delta < AD$ with probability at least $1 - \delta/(AD)$ we have*

$$\left|\hat{\mu}_a(T) - \mu_a\right| \leq \sqrt{\frac{1+N_a(T)}{N_a^2(T)}\left(1 + 2\log\left(\frac{AD\sqrt{1+N_a(T)}}{\delta}\right)\right)}, \quad \forall T \in \mathbb{N}.$$

Let $\mathrm{UC}_a^i := \cup_{t=1}^T \{\mu_a^i \notin [l_a^i(t), u_a^i(t)]\}$, $\mathrm{UC}^i := \cup_{a\in\mathcal{A}} \mathrm{UC}_a^i$ and $\mathrm{UC} := \cup_{i\in\mathcal{D}} \mathrm{UC}^i$. The following lemma bounds the probability of UC.

Lemma 2 $\mathbb{P}(\mathrm{UC}) \leq \delta$.

Proof This follows from Lemma 1. We observe that $\{\mu_a^i \in [l_a^i(t), u_a^i(t)]\} = \{|\mu_a^i - \hat{\mu}_a^i(t)| \leq c_a(t)\}$. Thus, Lemma 1 shows that $\neg\mathrm{UC}_a^i$ holds with probability at least $1 - \delta/(AD)$, and hence, UC_a^i holds with probability at most $\delta/(AD)$. Applying the union bound, we get $\mathbb{P}(\mathrm{UC}) \leq \delta$. □

Let $\mathcal{T} := \{1 \leq t \leq T : \forall a \in \hat{\mathcal{A}}_*^1(t) : c_a(t) \leq \epsilon/2\}$ denote the set of rounds in which PF-LEX selects the arm $\hat{a}_*^D(t)$ and $\neg\mathcal{T} := \{1, \ldots, T\} - \mathcal{T}$. In the following lemma, the gap of the arm selected in round $t \in \mathcal{T}$ in objective i is bounded as a function of ϵ and the length of the confidence interval of the selected arm on event $\neg$UC if the selected arm is in $\mathcal{S}_*^i$.

Lemma 3 *When PF-LEX is run, the following holds on event* $\neg$UC *if* $a(t) \in \mathcal{S}_*^i$: $\mu_*^i - \mu_{a(t)}^i \leq u_{a(t)}^i(t) - l_{a(t)}^i(t) + \epsilon(A-1)$ *for* $t \in \mathcal{T}$.

Proof Consider any lexicographic optimal arm $*$. We have

$$\begin{aligned} \mu_*^i - \mu_{a(t)}^i &\leq u_*^i(t) - l_{a(t)}^i(t) \\ &\leq u_{\hat{a}_*^i(t)}^i(t) - l_{a(t)}^i(t) \\ &\leq u_{a(t)}^i(t) - l_{a(t)}^i(t) + \epsilon(A-1). \end{aligned} \tag{1}$$

Here (1) holds since $\mu_*^i \leq u_*^i(t)$ and $\mu_{a(t)}^i \geq l_{a(t)}^i(t)$ on event $\neg$UC. Equation (1) holds by the definition of $\hat{a}_*^i(t)$ and the fact that $* \in \hat{\mathcal{A}}_*^{i-1}(t)$, which is proven by induction. For this, consider any objective $j \in \{1, \ldots, i-1\}$. We first observe that $* \in \hat{\mathcal{A}}_*^0(t)$ and $a(t) = \hat{a}_*^D(t) \in \hat{\mathcal{A}}_*^{D-1}(t) \subseteq \hat{\mathcal{A}}_*^j(t)$. Next, we show that $* \in \hat{\mathcal{A}}_*^{j-1}(t) \implies * \in \hat{\mathcal{A}}_*^j(t)$ to conclude that $* \in \hat{\mathcal{A}}_*^{i-1}(t)$. Since $a(t) \in \mathcal{S}_*^i$, $\mu_{a(t)}^j = \mu_*^j$, which implies that $a(t)$ and $*$ are linked in objective j. Since $a(t) \in \hat{\mathcal{A}}_*^j(t)$, $a(t)$ is chained to $\hat{a}_*^j(t)$ in objective j, which implies that $*$ is chained to $\hat{a}_*^j(t)$ in objective j as well. Finally, if $i = D$, (1) holds trivially as $a(t) = \hat{a}_*^i(t) = \hat{a}_*^D(t)$. Otherwise, since $a(t) = \hat{a}_*^D(t) \in \hat{\mathcal{A}}_*^{D-1}(t) \subseteq \hat{\mathcal{A}}_*^i(t)$, $a(t)$ is chained to $\hat{a}_*^i(t)$, which implies that $|u_{\hat{a}_*^i(t)}^i(t) - u_{a(t)}^i(t)| \leq 2(|\hat{\mathcal{A}}_*^i(t)| - 1)\max_{a\in\hat{\mathcal{A}}_*^i(t)} c_a(t) \leq \epsilon(A-1)$. □

We also need to bound the regret in objective i for rounds up to round T for which $t \notin \mathcal{T}$. Let $\neg\mathcal{T}_a := \{t \in \neg\mathcal{T} : a(t) = a\}$. Obviously, PF-LEX does not incur any regret in objective i in rounds $t \in \neg\mathcal{T}_a$ for $a \in \mathcal{A} - \mathcal{S}_*^i$, and incurs regret Δ_a^i in objective i in rounds $t \in \neg\mathcal{T}_a$ for $a \in \mathcal{S}_*^i$.

Lemma 4 *When PF-LEX is run, we have*

$$\sum_{t\in\neg\mathcal{T}} \mathbb{I}\{a(t)\in\mathcal{S}^i_*\}\Delta^i_{a(t)} \le \sum_{a\in\mathcal{S}^*_i}\left(3+\frac{16}{\epsilon^2}\log\frac{2\sqrt{e}AD}{\epsilon\delta}\right)\Delta^i_a$$

for all objectives $i\in\mathcal{D}$.

Proof The proof follows from bounding the cardinality of $\neg\mathcal{T}_a$ for $a\in\mathcal{S}^i_*$. Note that $t\in\neg\mathcal{T}_a$ happens only when $c_a(t)>\epsilon/2$. Similar to the proof of Theorem 7 in Abbasi-Yadkori et al., (2011), this implies that

$$\frac{N_a^2(t)-1}{N_a(t)+1}\le\frac{N_a^2(t)}{N_a(t)+1}\le\frac{4}{\epsilon^2}\left(1+2\log\frac{AD\sqrt{1+N_a(t)}}{\delta}\right).$$

Then, from Lemma 8 in Antos et al. (2010), we obtain $N_a(t)\le 3+\frac{16}{\epsilon^2}\log\frac{2\sqrt{e}AD}{\epsilon\delta}$. □

In the remaining part, we bound $\mathrm{Reg}^i_{pb}(T)$ under the event ¬UC and $\mathbb{E}[\mathrm{Reg}^i_{pb}(T)]$ by using the results of the lemmas above. For the latter, we observe that:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}^i_{pb}(T)] &= \mathbb{E}[\mathrm{Reg}^i_{pb}(T)|\mathrm{UC}]\mathbb{P}(\mathrm{UC})+\mathbb{E}[\mathrm{Reg}^i_{pb}(T)|\neg\mathrm{UC}]\mathbb{P}(\neg\mathrm{UC})\\ &\le T\Delta^i_{\max}\mathbb{P}(\mathrm{UC})+\mathbb{E}[\mathrm{Reg}^i_{pb}(T)|\neg\mathrm{UC}].\end{aligned}\tag{2}$$

For each $i\in\mathcal{D}$, the bound for $\mathrm{Reg}^i_{pb}(T)$ is obtained by using the result in Lemmas 3 and 4. By Lemma 4, we know that

$$\sum_{t\in\neg\mathcal{T}}\mathbb{I}\{a(t)\in\mathcal{S}^*_i\}\Delta^i_{a(t)}\le 3|\mathcal{S}^i_*|\Delta^i_{\max}+\frac{16|\mathcal{S}^i_*|\Delta^i_{\max}}{\epsilon^2}\log\frac{2\sqrt{e}AD}{\epsilon\delta}.\tag{3}$$

Let $\mathcal{N}_a:=\{t\in\mathcal{T}: a(t)=a\}$. By Lemma 3, on event ¬UC (which happens with probability at least $1-\delta$), we have

$$\begin{aligned}\sum_{t\in\mathcal{T}}\mathbb{I}\{a(t)\in\mathcal{S}^*_i\}\Delta^i_{a(t)} &\le \sum_{a\in\mathcal{S}^i_*}\sum_{t\in\mathcal{N}_a}(u^i_a(t)-l^i_a(t))+\epsilon(A-1)T\\ &\le 2\sqrt{2}\sum_{a\in\mathcal{S}^i_*}\left(B_{T,\delta}\sum_{t\in\mathcal{N}_a}\sqrt{\frac{1}{N_a(t)}}\right)+\epsilon(A-1)T\\ &\le 2\sqrt{2}B_{T,\delta}\sum_{a\in\mathcal{S}^i_*}\sqrt{N_a(T)}+\epsilon(A-1)T\\ &\le 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}^i_*|T}+\epsilon(A-1)T.\end{aligned}\tag{4}$$

The bound for $\mathrm{Reg}^i_{pb}(T)$ is obtained by summing the results of (3) and (4). Finally, the bounds on the expected regret simply follows from using (4) and setting $\delta=1/T$.

Regret lower bound for Case 1 The following theorem shows that the total priority-based regret of any algorithm is at least on the order of $\Omega(T^{2/3})$. Such a lower bound has two major implications: (i) the regret achieved by PF-LEX is optimal up to some logarithmic terms, (ii) the Lex-MAB is inherently a “harder” problem than conventional MAB problems, where achieving a gap-free regret of $O(T^{1/2})$ is usually possible.

Theorem 2 *Define* $\mathrm{Reg}_\Sigma(T) := \sum_{i=1}^{D} \mathrm{Reg}^i_{pb}(T)$ *as the total regret. For any algorithm and for any* $\epsilon > 0$, *there exists some instance of the Lex-MAB such that*

$$\mathbb{E}[\mathrm{Reg}_\Sigma(T)] \geq \min\left\{\frac{1}{\epsilon^2}, \frac{\epsilon T}{2} - \frac{1}{\epsilon}\right\}.$$

Taking $\epsilon = T^{-1/3}$, *for any algorithm, there exists some instance of the Lex-MAB such that* $\mathbb{E}[\mathrm{Reg}_\Sigma(T)] \geq \Omega(T^{2/3})$.

Proof of Theorem 2 For any given algorithm, consider two instances ϕ and ψ of the Lex-MAB with $A = 2$ and $D = 2$. An instance includes both the probabilistic structure of the given learning algorithm and an environment consisting of arms. Arms in instance ϕ has expected reward vectors $\boldsymbol{\mu}_1^{(\phi)} = (1, 0)$ and $\boldsymbol{\mu}_2^{(\phi)} = (1, 1)$ while arms in instance ψ has expected reward vectors $\boldsymbol{\mu}_1^{(\psi)} = (1 + \varepsilon, 0)$ and $\boldsymbol{\mu}_2^{(\psi)} = (1, 1)$. We assume rewards are distributed independently for each objective and normally with unit variance in both instances.

Then, the expected total regret of an algorithm in instances ϕ and ψ can be written as

$$\mathbb{E}_\phi[\mathrm{Reg}_\Sigma(T)] = \mathbb{E}_\phi[N_1(T)], \tag{5}$$

$$\begin{aligned}\mathbb{E}_\psi[\mathrm{Reg}_\Sigma(T)] &= \mathbb{E}_\psi[N_2(T)]\varepsilon \\ &= \varepsilon(T - \mathbb{E}_\psi[N_1(T)])\end{aligned} \tag{6}$$

respectively. If $\mathbb{E}_\phi[N_1(T)] \geq 1/\varepsilon^2$, then the regret in (5) is simply lower bounded as $\mathbb{E}_\phi[\mathrm{Reg}_\Sigma(T)] = \mathbb{E}_\phi[N_1(T)] \geq 1/\varepsilon^2$.

If $\mathbb{E}_\phi[N_1(T)] \leq 1/\varepsilon^2$ instead, then we first apply inequality (6) in Garivier et al. (2018), which trivially extends to multivariate distributions, to obtain

$$\begin{aligned}\mathbb{E}_\phi[N_1(T)] \cdot \frac{\varepsilon^2}{2} &= \mathbb{E}_\phi[N_1(T)] KL(\mathcal{N}(\boldsymbol{\mu}_1^{(\phi)}, \boldsymbol{I}), \mathcal{N}(\boldsymbol{\mu}_1^{(\psi)}, \boldsymbol{I})) \\ &\geq kl\left(\frac{\mathbb{E}_\phi[N_1(T)]}{T}, \frac{\mathbb{E}_\psi[N_1(T)]}{T}\right) \\ &\geq 2\left(\frac{\mathbb{E}_\phi[N_1(T)]}{T} - \frac{\mathbb{E}_\psi[N_1(T)]}{T}\right)^2\end{aligned}$$

where $KL(v, v')$ represents the Kullback–Leibler divergence between distributions v and v', $kl(p, q) := p\ln(p/q) + (1-p)\ln((1-p)/(1-q))$, and the last line follows from Pinsker's inequality. Solving for $\mathbb{E}_\psi[N_1(T)]/T$, we obtain

$$\begin{aligned}\frac{\mathbb{E}_\psi[N_1(T)]}{T} &\leq \frac{\mathbb{E}_\phi[N_1(T)]}{T} + \frac{\varepsilon}{2}\sqrt{\mathbb{E}_\phi[N_1(T)]} \\ &\leq \frac{1}{\varepsilon^2 T} + \frac{1}{2} \\ \implies \quad \mathbb{E}_\psi[N_1(T)] &\leq \frac{1}{\varepsilon^2} + \frac{T}{2}.\end{aligned}$$

Then, the regret in (6) can be lower bounded as

$$\begin{aligned}
\mathbb{E}_{\psi}[\mathrm{Reg}_{\Sigma}(T)] &= \varepsilon(T - \mathbb{E}_{\psi}[N_1(T)]) \\
&\geq \varepsilon\left(T - \frac{1}{\varepsilon^2} - \frac{T}{2}\right) \\
&= \frac{\varepsilon T}{2} - \frac{1}{\varepsilon}.
\end{aligned}$$

In all cases, we have either lower bounded the regret in instance ψ or instance ϕ. Combining those lower bounds, there is at least one instance where the expected total regret is lower bounded as

$$\mathbb{E}[\mathrm{Reg}_{\Sigma}(T)] \geq \min\left\{\frac{1}{\varepsilon^2}, \frac{\varepsilon T}{2} - \frac{1}{\varepsilon}\right\}.$$

□

A learning algorithm for Case 2 We propose *Optimal Mean based Lexicographic Exploration and eXploitation* (OM-LEX) given in Algorithm 2 for the prior information described in Case 2. In essence, OM-LEX generalizes the arm selection rule proposed in Algorithm 1 in Garivier et al. (2018) to multiple objectives. Similar to PF-LEX, it keeps, for each arm a, the counter $N_a(t)$ and the sample mean reward $\hat{\mu}^i_a(t)$, $\forall i \in \mathcal{D}$.

Algorithm 2 OM-LEX

```
**Inputs:** $\mu^i_* = 0, \forall i \in \mathcal{D}$
**Counters:** $N_a, \forall a \in \mathcal{A}$
**Estimates:** $\hat{\mu}^i_a, \forall a \in \mathcal{A}, \forall i \in \mathcal{D}$
For each round $t \in \{1, \ldots, A\}$, select arm $t$
For each round $t > A$:
    Compute $\hat{\mathcal{A}}_* = \{a \in \mathcal{A} : \forall i \in \mathcal{D}, |\hat{\mu}^i_a| < \sqrt{4 \log N_a / N_a}\}$
    If $\hat{\mathcal{A}}_* \neq \emptyset$, select an arm $a(t)$ in $\hat{\mathcal{A}}_*$ uniformly at random, update $t \leftarrow t+1$
    If $\hat{\mathcal{A}}_* = \emptyset$, select $a(t) = 1, a(t+1) = 2, \ldots, a(t+A-1) = A$, update $t \leftarrow t+A$
```

OM-LEX starts by selecting each arm exactly once. In the remaining rounds, it checks whether there exists an arm whose sample mean reward in objective i is within a shrinking neighborhood of the lexicographic optimal arm's expected reward for all objectives $i \in \mathcal{D}$. For this, it computes the set of estimated lexicographic optimal arms in round t as

$$\hat{\mathcal{A}}_*(t) := \left\{a \in \mathcal{A} : \forall i \in \mathcal{D}, |\hat{\mu}^i_a(t)| < \sqrt{\frac{4 \log N_a(t)}{N_a(t)}}\right\}.$$

If $\hat{\mathcal{A}}_*(t) \neq \emptyset$, then OM-LEX exploits by selecting one of the arms in $\hat{\mathcal{A}}_*(t)$ uniformly at random as it expects only the lexicographic optimal arms to satisfy this condition in the long run. If no such arm exists, then OM-LEX explores by playing all arms in a round-robin fashion. The following theorem shows that the expected priority-based regret of OM-LEX is uniformly bounded in time.

Theorem 3 *When OM-LEX is run*, $\forall i \in \mathcal{D}$ *and* $\forall T \geq 1$, *we have*

$$\mathbb{E}[\mathrm{Reg}^i_{pb}(T)] \leq \sum_{a \in \mathcal{S}^i_*} \left(\left(\frac{\pi^2}{3} D + 1 \right) \Delta^i_a + \frac{36 \Delta^i_a}{(\nabla^{\max}_a)^2} \log \frac{17}{\nabla^{\max}_a} \right).$$

When $D = 1$, this result is identical to the regret bound in Theorem 9 in Garivier et al. (2018) except for some constants. In the multi-objective case, we see that the regret induced by an arm in one objective depends on the maximum of the absolute gaps of the same arm over all objectives. As long as the arm has a large absolute gap in at least one objective, it is easy to identify it as a suboptimal arm.

Proof of Theorem 3 We use the following fact to prove Theorem 3 (and Theorem 5 later on).

Fact 1 (Results from the proof of Theorem 9 in Garivier et al., 2018) *Given* $\Delta > 0$, *for all arms* $a \in \mathcal{A}$, *for all objectives* $i \in \mathcal{D}$ *and for all rounds* $t \in \{1, 2, \dots, T\}$, *we have*

$$\begin{aligned}
&\sum_{w=1}^{\infty} \mathbb{P}\left(\hat{\mu}^i_a(t) - \mu^i_a > \Delta - \sqrt{\frac{4 \log N_a(t)}{N_a(t)}} \,\middle|\, N_a(t) = w \right) \\
&= \sum_{w=1}^{\infty} \mathbb{P}\left(\hat{\mu}^i_a(t) - \mu^i_a < \sqrt{\frac{4 \log N_a(t)}{N_a(t)}} - \Delta \,\middle|\, N_a(t) = w \right) \\
&\leq \frac{36}{\Delta^2} \log \frac{17}{\Delta}.
\end{aligned}$$

For an arm a that is not lexicographic optimal, let $\dagger(a) := \mathrm{argmax}_{j \in \mathcal{D}} \nabla^j_a$ so that $\nabla^{\dagger(a)}_a = \nabla^{\max}_a$. When a can be inferred from the context, $\dagger(a)$ is denoted by $\dagger$ only. For all objectives $i \in \mathcal{D}$, we decompose $\mathbb{E}[\mathrm{Reg}^i(T)]$ as

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}^i(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{a(t) \in \mathcal{S}^i_*\} \Delta^i_{a(t)} \right] \\
&= \mathbb{E}\left[\sum_{a \in \mathcal{S}^i_*} \sum_{t=1}^{T} \mathbb{I}\{a(t) = a\} \Delta^i_a \right] \qquad (7) \\
&= \mathbb{E}\left[\sum_{a \in \mathcal{S}^i_*} \sum_{t=1}^{T} \mathbb{I}\{t \leq A, a(t) = a\} \Delta^i_a \right]
\end{aligned}$$

$$+ \mathbb{E}\left[\sum_{a \in \mathcal{S}^i_*} \sum_{t=1}^{T} \mathbb{I}\left\{ t > A, |\hat{\mu}^{\dagger}_a(t)| < \sqrt{\frac{4 \log N_a(t)}{N_a(t)}}, a(t) = a \right\} \Delta^i_a \right] \qquad (8)$$

$$+\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{t>A, |\hat{\mu}^\dagger_a(t)|\geq\sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t)=a\right\}\Delta^i_a\right]. \tag{9}$$

Bounding (7) is trivial, since each arm is played exactly once for rounds $t \leq A$. We have

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\{t\leq A, a(t)=a\}\Delta^i_a\right] &= \mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\{t=a\}\Delta^i_a\right]\\ &=\sum_{a\in\mathcal{S}^i_*}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{t=a\}\right]\Delta^i_a\\ &=\sum_{a\in\mathcal{S}^i_*}\Delta^i_a.\end{aligned}$$

In order to bound (8), we define $\tau_a(w)$ as the wth round for which $a(t)=a$, and $w_a(T)$ as the number of rounds for which $a(t)=a$ by round T. By definition, $N_a(\tau_a(w+1))=w$, $\tau_a(1)\leq A$ and $w_a(T)\leq T$ hold for all arms a. Thus, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{t>A, |\hat{\mu}^\dagger_a(t)|<\sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t)=a\right\}\Delta^i_a\right]\\ &=\sum_{a\in\mathcal{S}^i_*}\mathbb{E}\left[\sum_{w=1}^{w_a(T)-1}\sum_{t=\tau_a(w)+1}^{\tau_a(w+1)}\mathbb{I}\left\{|\hat{\mu}^\dagger_a(t)|<\sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t)=a\right\}\right]\Delta^i_a\\ &=\sum_{a\in\mathcal{S}^i_*}\mathbb{E}\left[\sum_{w=1}^{w_a(T)-1}\mathbb{I}\left\{|\hat{\mu}^\dagger_a(\tau_a(w+1))|<\sqrt{\frac{4\log w}{w}}\right\}\right]\Delta^i_a\\ &\leq\sum_{a\in\mathcal{S}^i_*}\sum_{w=1}^{\infty}\mathbb{P}\left(|\hat{\mu}^\dagger_a(\tau_a(w+1))|<\sqrt{\frac{4\log w}{w}}\right)\Delta^i_a.\end{aligned} \tag{10}$$

When $\mu^\dagger_a=\nabla^\dagger_a$, we have

$$\begin{aligned}&\sum_{w=1}^{\infty}\mathbb{P}\left(|\hat{\mu}^\dagger_a(\tau_a(w+1))|<\sqrt{\frac{4\log w}{w}}\right)\\ &\leq\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^\dagger_a(\tau_a(w+1))<\sqrt{\frac{4\log w}{w}}\right)\\ &\leq\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^\dagger_a(\tau_a(w+1))-\mu^\dagger_a<\sqrt{\frac{4\log w}{w}}-\nabla^\dagger_a\right)\\ &\leq\frac{36}{(\nabla^\dagger_a)^2}\log\frac{17}{\nabla^\dagger_a},\end{aligned} \tag{11}$$

where (11) is due to Fact 1.

Similarly, when $\mu_a^\dagger = -\nabla_a^\dagger$, we have

$$\begin{aligned}\sum_{w=1}^{\infty} \mathbb{P}\left(|\hat{\mu}_a^\dagger(\tau_a(w+1))| < \sqrt{\frac{4\log w}{w}} \right) \\ &\leq \sum_{w=1}^{\infty} \mathbb{P}\left(\hat{\mu}_a^\dagger(\tau_a(w+1)) > -\sqrt{\frac{4\log w}{w}} \right) \\ &\leq \sum_{w=1}^{\infty} \mathbb{P}\left(\hat{\mu}_a^\dagger(\tau_a(w+1)) - \mu_a^\dagger > \nabla_a^\dagger - \sqrt{\frac{4\log w}{w}} \right) \\ &\leq \frac{36}{(\nabla_a^\dagger)^2} \log \frac{17}{\nabla_a^\dagger},\end{aligned} \tag{12}$$

where (12) is again due to Fact 1.

Combining (11) and (12), we obtain

$$(12) \leq \sum_{a \in \mathcal{S}_*^i} \frac{36\Delta_a^i}{(\nabla_a^\dagger)^2} \log \frac{17}{\nabla_a^\dagger}.$$

In order to bound (9), we observe that $t > A \wedge |\hat{\mu}_a^\dagger(t)| \geq \sqrt{4\log N_a(t)/N_a(t)} \wedge a(t) = a$ can only occur during an exploration stage, where each arm is played successively. Hence, we can infer that

(i) $a(t - a + *) = *$,
(ii) $t - a + 1 > A$,
(iii) $\hat{\mathcal{A}}_*(t - a + 1) = \emptyset$, which implies that there exists an objective j such that

$$\begin{aligned}|\hat{\mu}_*^j(t - a + *)| = |\hat{\mu}_*^j(t - a + 1)| &\geq \sqrt{4\frac{\log N_*(t-a+1)}{N_*(t-a+1)}} \\ &= \sqrt{4\frac{\log N_*(t-a+*)}{N_*(t-a+*)}},\end{aligned}$$

since arm $*$ is not played after round $t - a + 1$ until round $t - a + *$.

Using these observations and defining $t_a := t - a + *$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{t > A, |\hat{\mu}^{\dagger}_a(t)| \geq \sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t) = a\right\}\Delta^i_a\right]\\
&= \mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{\exists j : t_a > A, |\hat{\mu}^{j}_*(t_a)| \geq \sqrt{\frac{4\log N_*(t_a)}{N_*(t_a)}}, a(t_a) =*\right\}\Delta^i_a\right]\\
&\leq \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{t_a > A, |\hat{\mu}^{j}_*(t_a)| \geq \sqrt{\frac{4\log N_*(t_a)}{N_*(t_a)}}, a(t_a) =*\right\}\Delta^i_a\right]\\
&\leq \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{t > A, |\hat{\mu}^{j}_*(t)| \geq \sqrt{\frac{4\log N_*(t)}{N_*(t)}}, a(t) =*\right\}\Delta^i_a\right]
\end{aligned}
$$

$$
= \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{w=1}^{w_*(T)-1}\sum_{t=\tau_*(w)+1}^{\tau_*(w+1)}\mathbb{I}\left\{|\hat{\mu}^{j}_*(t)| \geq \sqrt{\frac{4\log N_*(t)}{N_*(t)}}, a(t) =*\right\}\right]\Delta^i_a \tag{13}
$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{w=1}^{w_*(T)-1}\mathbb{I}\left\{|\hat{\mu}^{j}_*(\tau_*(w+1))| \geq \sqrt{\frac{4\log w}{w}}\right\}\right]\Delta^i_a\\
&\leq \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\mathbb{P}\left(|\hat{\mu}^{j}_*(\tau_*(w+1))| \geq \sqrt{\frac{4\log w}{w}}\right)\Delta^i_a\\
&= \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\left[\mathbb{P}\left(\hat{\mu}^{j}_*(\tau_*(w+1)) \leq -\sqrt{\frac{4\log w}{w}}\right)\right.\\
&\quad \left.+\ \mathbb{P}\left(\hat{\mu}^{j}_*(\tau_*(w+1)) \geq \sqrt{\frac{4\log w}{w}}\right)\right]\Delta^i_a\\
&\leq \sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\frac{2}{w^2}\Delta^i_a\\
&\leq \sum_{a\in\mathcal{S}^i_*}\left(\frac{\pi^2}{3}\right)D\Delta^i_a,
\end{aligned}
$$

where (13) is due to Hoeffding's inequality for sub-Gaussian random variables (Bubeck et al., 2013). □

Priority-free regret bound for Case 2 The following theorem shows that the expected priority-free regret of OM-LEX is uniformly bounded in time as well.

Theorem 4 *When OM-LEX is run*, $\forall i \in \mathcal{D}$ *and* $\forall T \geq 1$, *we have*

$$\mathbb{E}[\mathrm{Reg}^i_{pf}(T)] \leq \sum_{a \in \mathcal{S}^i} \left(\left(\frac{\pi^2}{3} D + 1 \right) \Delta^i_a + \frac{36}{\nabla^{\max}_a} \log \frac{17}{\nabla^{\max}_a} \right).$$

Proof Note that

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}^i_{pf}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \Delta^i_{a(t)} \right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} \Delta^i_{a(t)} \mathbb{I}\{a(t) \in \mathcal{S}^i\} \right] \qquad (14) \\ &\leq \sum_{a \in \mathcal{S}^i} \left(\left(\frac{\pi^2}{3} D + 1 \right) \Delta^i_a + \frac{36}{\nabla^{\max}_a} \log \frac{17}{\nabla^{\max}_a} \right) \end{aligned}$$

where we prove (14) by replacing $\mathcal{S}^i_*$ with $\mathcal{S}^i$ in the proof of Theorem 3. □

A learning algorithm for Case 3 We propose *Near Optimal Mean based Lexicographic Exploration and eXploitation* (NOM-LEX). NOM-LEX has almost the same structure with OM-LEX. Its pseudocode is exactly the same as Algorithm 2 except two differences: Firstly, its input prior knowledge (given in line 1 of Algorithm 2) is $\eta_i = 0$, $\forall i \in \mathcal{D}$. Secondly, NOM-LEX computes the set of estimated lexicographic optimal arms in round t (given in line 6 of Algorithm 2) as

$$\hat{\mathcal{A}}_*(t) := \left\{ a \in \mathcal{A} : \forall i \in \mathcal{D}, \hat{\mu}^i_a(t) > -\sqrt{\frac{4 \log N_a(t)}{N_a(t)}} \right\}.$$

The next theorem bounds the expected priority-based regret of NOM-LEX.

Theorem 5 *When NOM-LEX is run*, $\forall i \in \mathcal{D}$ *and* $\forall T \geq 1$, *we have*

$$\begin{aligned} &\mathbb{E}[\mathrm{Reg}^i_{pb}(T)] \\ &\leq \sum_{a \in \mathcal{S}^i_*} \left(\left(\frac{\pi^2}{6} D + 1 \right) \Delta^i_a + \frac{36 \Delta^i_a}{(\max_{j \in \mathcal{D}}(\Delta^j_a - \delta_j))^2} \log \frac{17}{\max_{j \in \mathcal{D}}(\Delta^j_a - \delta_j)} \right). \end{aligned}$$

From Theorem 5, we see that the regret due to a suboptimal arm a in objective i depends on the maximum squared difference between the suboptimality gaps of that arm and near-lexicographic optimal expected rewards over all objectives. This also shows that the prior knowledge in other objectives may help the learner attain smaller regret in objective i. However, since the lexicographic optimal expected rewards are not known, unlike Case 2, we cannot rule out a suboptimal arm in objective i by observing that it is much better than a lexicographic optimal arm in another objective.

Proof of Theorem 5 For an arm a that is not lexicographic optimal, let $\dagger(a) := \mathrm{argmax}_{j \in \mathcal{D}} \Delta^j_a - \delta_j$. When a can be inferred from the context, $\dagger(a)$ is denoted by $\dagger$ only. For all objectives $i \in \mathcal{D}$, we decompose $\mathbb{E}[\mathrm{Reg}^i(T)]$ as

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}^i(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{a(t) \in \mathcal{S}_*^i\} \varDelta_{a(t)}^i\right] \\ &= \mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\{a(t) = a\} \varDelta_a^i\right] \\ &= \mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\{t \le A, a(t) = a\} \varDelta_a^i\right] \end{aligned} \tag{15}$$

$$+ \mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\left\{t > A, \hat{\mu}_a^\dagger(t) > -\sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t) = a\right\}\varDelta_a^i\right] \tag{16}$$

$$+ \mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\left\{t > A, \hat{\mu}_a^\dagger(t) \le -\sqrt{\frac{4\log N_a(t)}{N_a(t)}}, a(t) = a\right\}\varDelta_a^i\right]. \tag{17}$$

Bounding (15) is trivial, since each arm is played exactly once for rounds $t \le A$. We have

$$\begin{aligned}\mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\{t \le A, a(t) = a\} \varDelta_a^i\right] &= \mathbb{E}\left[\sum_{a\in\mathcal{S}_*^i} \sum_{t=1}^{T} \mathbb{I}\{t = a\} \varDelta_a^i\right] \\ &= \sum_{a\in\mathcal{S}_*^i} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{t = a\}\right] \varDelta_a^i \\ &= \sum_{a\in\mathcal{S}_*^i} \varDelta_a^i. \end{aligned}$$

In order to bound (16), we use $\tau_a(w)$ and $w_a(T)$ defined in the proof of Theorem 3. We have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{t>A,\hat{\mu}^\dagger_a(t)>-\sqrt{\frac{4\log N_a(t)}{N_a(t)}},a(t)=a\right\}\Delta^i_a\right]\\
&=\sum_{a\in\mathcal{S}^i_*}\mathbb{E}\left[\sum_{w=1}^{w_a(T)-1}\sum_{t=\tau_a(w)+1}^{\tau_a(w+1)}\mathbb{I}\left\{\hat{\mu}^\dagger_a(t)>-\sqrt{\frac{4\log N_a(t)}{N_a(t)}},a(t)=a\right\}\right]\Delta^i_a\\
&=\sum_{a\in\mathcal{S}^i_*}\mathbb{E}\left[\sum_{w=1}^{w_a(T)-1}\mathbb{I}\left\{\hat{\mu}^\dagger_a(\tau_a(w+1))>-\sqrt{\frac{4\log w}{w}}\right\}\right]\Delta^i_a \qquad (18)\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^\dagger_a(\tau_a(w+1))>-\sqrt{\frac{4\log w}{w}}\right)\Delta^i_a\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^\dagger_a(\tau_a(w+1))-\mu^\dagger_a>\Delta^\dagger_a-\delta_{\dagger(a)}-\sqrt{\frac{4\log w}{w}}\right)\Delta^i_a\\
&\le\sum_{a\in\mathcal{S}^i_*}\frac{36\Delta^i_a}{(\Delta^\dagger_a-\delta_{\dagger(a)})^2}\log\frac{17}{\Delta^\dagger_a-\delta_{\dagger(a)}},
\end{aligned}
$$

where (18) is due to Fact 1.

In order to bound (17), we observe that $t>A\wedge\hat{\mu}^\dagger_a(t)\le-\sqrt{4\log N_a(t)/N_a(t)}\wedge a(t)=a$ can only occur during an exploration stage, where each arm is played successively. Hence we can infer that

(i) $a(t-a+*)=*$,
(ii) $t-a+1>A$,
(iii) $\hat{\mathcal{A}}_*(t-a+1)=\emptyset$, which implies that there exists an objective j such that

$$
\begin{aligned}
\hat{\mu}^j_*(t-a+*)=\hat{\mu}^j_*(t-a+1)&\le-\sqrt{4\frac{\log N_*(t-a+1)}{N_*(t-a+1)}}\\
&=-\sqrt{4\frac{\log N_*(t-a+*)}{N_*(t-a+*)}},
\end{aligned}
$$

since arm $*$ is not played after round $t-a+1$ until round $t-a+*$.

Using these observations and defining $t_a:=t-a+*$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{t>A,\hat{\mu}^{\dagger}_{a}(t)\le-\sqrt{\frac{4\log N_a(t)}{N_a(t)}},a(t)=a\right\}\Delta^i_a\right]\\
&=\mathbb{E}\left[\sum_{a\in\mathcal{S}^i_*}\sum_{t=1}^{T}\mathbb{I}\left\{\exists j: t_a>A,\hat{\mu}^{j}_{*}(t_a)\le-\sqrt{\frac{4\log N_*(t_a)}{N_*(t_a)}},a(t_a)=*\right\}\Delta^i_a\right]\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{t_a>A,\hat{\mu}^{j}_{*}(t_a)\le-\sqrt{\frac{4\log N_*(t_a)}{N_*(t_a)}},a(t_a)=*\right\}\Delta^i_a\right]\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{t>A,\hat{\mu}^{j}_{*}(t)\le-\sqrt{\frac{4\log N_*(t)}{N_*(t)}},a(t)=*\right\}\Delta^i_a\right]
\end{aligned}\tag{19}$$

$$\begin{aligned}
&=\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{w=1}^{w_*(T)-1}\sum_{t=\tau_*(w)+1}^{\tau_*(w+1)}\mathbb{I}\left\{\hat{\mu}^{j}_{*}(t)\le-\sqrt{\frac{4\log N_*(t)}{N_*(t)}},a(t)=*\right\}\right]\Delta^i_a\\
&=\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\mathbb{E}\left[\sum_{w=1}^{w_*(T)-1}\mathbb{I}\left\{\hat{\mu}^{j}_{*}(\tau_*(w+1))\le-\sqrt{\frac{4\log w}{w}}\right\}\right]\Delta^i_a\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^{j}_{*}(\tau_*(w+1))\le-\sqrt{\frac{4\log w}{w}}\right)\Delta^i_a\\
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\mathbb{P}\left(\hat{\mu}^{j}_{*}(\tau_*(w+1))-\mu^j_*\le-\sqrt{\frac{4\log w}{w}}\right)\Delta^i_a
\end{aligned}$$

$$\begin{aligned}
&\le\sum_{a\in\mathcal{S}^i_*}\sum_{j=1}^{D}\sum_{w=1}^{\infty}\frac{1}{w^2}\Delta^i_a\\
&\le\sum_{a\in\mathcal{S}^i_*}\left(\frac{\pi^2}{6}\right)D\Delta^i_a,
\end{aligned}\tag{20}$$

where (19) holds since $\mu^j_*=\delta_j\ge0$ and (20) is due to Hoeffding's inequality for sub-Gaussian random variables (Bubeck et al., 2013). □

Priority-free regret bound for Case 3 The following theorem shows that the expected priority-free regret of NOM-LEX is uniformly bounded in time as well.

Theorem 6 *Redefine* $\Delta^i_{\min}:=\min_{a\in\mathcal{S}^i}\Delta^i_a$, $\forall i\in\mathcal{D}$. *When NOM-LEX is run*, $\forall i\in\mathcal{D}$ *and* $\forall T\ge1$, *we have*

$$\mathbb{E}[\mathrm{Reg}^i_{pf}(T)] \le \sum_{a\in\mathcal{S}^i}\left(\left(\frac{\pi^2}{6}D+1\right)\Delta^i_a + \frac{36\Delta^i_a}{(\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j))^2}\log\frac{17}{\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j)}\right).$$

Proof We redefine $\Delta^i_{\min}$ as stated in the theorem. Then,

$$\mathbb{E}[\mathrm{Reg}^i_{pf}(T)] \tag{21}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\Delta^i_{a(t)}\right]$$

$$\le \mathbb{E}\left[\sum_{t=1}^{T}\Delta^i_{a(t)}\mathbb{I}\{a(t)\in\mathcal{S}^i\}\right] \tag{22}$$

$$\le \sum_{a\in\mathcal{S}^i}\left(\left(\frac{\pi^2}{6}D+1\right)\Delta^i_a + \frac{36\Delta^i_a}{(\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j))^2}\log\frac{17}{\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j)}\right)$$

where we prove (22) by replacing $\mathcal{S}^i_*$ with $\mathcal{S}^i$ in the proof of Theorem 5. □

Note that redefining $\Delta^i_{\min}$ as $\min_{a\in\mathcal{S}^i}\Delta^i_a$ in Case 3 implies that the learner has stronger prior knowledge on the near-lexicographic optimal expected rewards, since $\min_{a\in\mathcal{S}^i}\Delta^i_a \le \min_{a\in\mathcal{S}^i_*}\Delta^i_a$.

4.2 Algorithms and regret bounds for the Sat-MAB

Assuming $\eta_i = 0$ for all $i\in\mathcal{D}$ without any loss of generality,[7] the algorithm proposed for Case 3, which is NOM-LEX, can also be used to solve the Sat-MAB. Since the goal now is to minimize the satisficing regret rather than the lexicographic regret, we no longer need η_i to lie between the lexicographic optimal and the second highest lexicographic optimal expected rewards in objective i. The following theorem bounds the expected satisficing regret for NOM-LEX.

Theorem 7 *When NOM-LEX is run for the satisficing goal*, $\forall i\in\mathcal{D}$ *and* $\forall T\ge 1$, *we have*

$$\mathbb{E}[\mathrm{Reg}^i_s(T)] \le \sum_{a\in\mathcal{S}^i_s}\left(\left(\frac{\pi^2}{6}D+1\right)(\Delta^i_a-\delta_i) + \frac{36(\Delta^j_a-\delta_j)}{(\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j))^2}\log\frac{17}{\max_{j\in\mathcal{D}}(\Delta^j_a-\delta_j)}\right).$$

Proof Replacing lexicographic optimality with satisficing optimality, $\mathcal{S}^i_*$ with $\mathcal{S}^i_s$, and every instance of the exact phrase Δ^i_a (and $\Delta^i_{a(t)}$) with $\Delta^i_a-\delta_i$ (and $\Delta^i_{a(t)}-\delta_i$), the proof of Theorem 5 holds for Theorem 7 as well. □

[7] See Remark 1 for a detailed discussion of why the generality is not lost.

Table 1 Expected reward vectors for the first three settings

Setting	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$
Setting 1	(0.50, 0.50)	(0.50, 0.40)	(0.40, 0.90)
Setting 2	(0.50, 0.50)	(0.50, 0.40)	(0.40, 0.50)
Setting 3	(0.50, 0.50)	(0.50, 0.40)	(0.40, 0.10)

Corollary 1 *When NOM-LEX is run for the single-objective satisfaction-in-mean-reward problem* $(D = 1)$, $\forall T \geq 1$, *we have*[8]

$$\mathbb{E}[\mathrm{Reg}_s(T)] \leq \sum_{a \in \mathcal{S}_s} \left(\left(\frac{\pi^2}{6} D + 1 \right) (\Delta_a - \delta) + \frac{36}{\Delta_a - \delta} \log \frac{17}{\Delta_a - \delta} \right).$$

Remark 4 Theorem 7 and Corollary 1 show that bounded regret is possible for the satisfaction-in-mean-reward problem. This result is directly in conflict with Corollary 2 of Reverdy et al. (2017), which claims a logarithmic lower bound on the single-objective case, and suggests that satisfaction-in-mean-reward UCL algorithm given in Section VI-A of Reverdy et al. (2017) is not optimal since it fails to achieve bounded regret.

5 Experiments

5.1 Experiments for the Lex-MAB

In this section, we demonstrate our results for the Lex-MAB in three different settings with $A = 3$ and $D = 2$ and two additional settings with $D = 3$. All rewards are assumed to come from independent Bernoulli distributions in all objectives.

For the first three settings with $A = 3$ and $D = 2$, the expected reward vectors are summarized in Table 1. In all of these settings, the only lexicographic optimal arm is the first arm and $\Delta^1_{\min} = \Delta^2_{\min} = 0.10$. Note that we only focus on the priority-based regret for these settings. In Setting 1, apart from the lexicographic optimal arm, there is another arm that is also optimal in objective 1, which requires the learner to consider rewards in objective 2. However, the third arm makes this tricky. It is not only suboptimal in objective 1 but also has a very high reward in objective 2. Setting 2 is specifically designed to be challenging for Cases 2 and 3. Since arms that are not lexicographic optimal are suboptimal in exactly one objective, eliminating arms based on information from the other objective is not possible. Setting 3 contrasts with Setting 1. Unlike Setting 1, in which the expected reward of arm 3 in objective is much higher than the lexicographic optimal expected reward, in Setting 3, it is much lower. However, the gap of arm 3 in objective 2 in Setting 3 is still the same as the absolute gap of arm 3 in Setting 1.

For all cases, we set $T = 10^5$ and average the regret of the learners over 100 individual runs. We consider OM-LEX, NOM-LEX, and PF-LEX with prior knowledge and parameters that are summarized in Table 2.[9] For PF-LEX, we do not consider the choices for

[8] For simplicity, the objective index 1 is omitted.

[9] Implementations of OM-LEX, NOM-LEX, and PF-LEX that are used during the experiments can be found at https://github.com/Bilkent-CYBORG/Lex-MAB.

Table 2 Prior knowledge and parameters of algorithms for the first three settings

Algorithm	Prior knowledge and parameters
OM-LEX (OM) 1	$\mu_*^1 = \mu_*^2 = 0.50$
NOM-LEX (NM) 1	$\eta_1 = \eta_2 = 0.45$
NOM-LEX (NM) 2	$\eta_1 = \eta_2 = 0.40 + 10^{-6}$
NOM-LEX (NM) 3	$\eta_1 = \eta_2 = 0.50 - 10^{-6}$
PF-LEX (PF) 1	$\epsilon = \delta = T^{-1/5}$
PF-LEX (PF) 2	$\epsilon = \delta = T^{-1/10}$
PF-LEX (PF) 3	$\epsilon = \delta = T^{-1/3}$

ϵ and δ given in Theorem 1 because they require a large number of rounds for the initial exploration stage of the algorithm. Instead, we consider different exponents of T as both ϵ and δ, except for a single result which shows the regret of PF-LEX for $\epsilon = \delta = T^{-1/3}$ (PF-LEX 3) for $T = 5 \times 10^8$.

Table 3 shows the regrets incurred and the percentage of rounds where a lexicographic optimal arm has been played by OM-LEX 1, NOM-LEX 1, 2, 3, and PF-LEX 1, 2 in Settings 1, 2 and 3 at $T = 10^5$. There, we also report the performance of the variants of OM-LEX and NOM-LEX that only learn from the first objective and ignore the second objective, i.e., they act as if $D = 1$. Note that Settings 1, 2, and 3 are equivalent for objective 1.

By looking at the regrets in objective 1 of OM-LEX 1, NOM-LEX 1, and their single-objective variants, we observe how information from objective 2 helps learning in objective 1. OM-LEX takes advantage of large absolute gaps independent from whether the actual mean reward is higher or lower than the mean reward of arm 1. As a result, in Settings 1 and 3, it achieves lower regret in objective 1 than its single-objective variant does. NOM-LEX is capable of doing this only when the gap is positive, a large absolute gap is not sufficient. As a result, only in Setting 3, it outperforms its single-objective variant. In Setting 2, where information from objective 2 is not as useful as it is in Settings 1 and 3 to rule out the suboptimal arm in objective 1, OM-LEX 1 performs worse than the other settings in objective 1.

By looking at the regrets of NOM-LEX 1, 2, and 3, we observe how different prior information affects the performance of NOM-LEX. Consistent with the proven regret bounds, knowing near optimal expected rewards that are closer to the lexicographic optimal ones decreases the regret in all objectives. When the near optimal expected rewards are extremely close to the lexicographic optimal ones, the performance of NOM-LEX is very similar to that of OM-LEX.

By looking at the percentage of rounds where a lexicographic optimal arms have been played, we see that the single-objective variants do not play lexicographic optimal arms as often as their multiobjective counterparts although they achive lower regret in objective 1 in some settings compared to the multiobjective variants.

Figure 1 shows the regrets of OM-LEX 1, NOM-LEX 1, PF-LEX 1 in Setting 1. We observe that the regret of OM-LEX in objective 1 is significantly smaller than the regret of NOM-LEX. We believe this is the case because OM-LEX is able to take advantage of the large absolute gap of arm 3 to eliminate it early on, whereas NOM-LEX cannot. The behavior of PF-LEX is explained as follows. Until around round 30,000, it explores all three arms uniformly since their estimated rewards in objective 1 are still chained to each other. At round 30,000, the gap between the arms is deemed small enough with respect

Table 3 Regrets incurred and the percentage of rounds where a lexicographic optimal arm has been played by OM-LEX 1, NOM-LEX 1–3, and PF-LEX 1–2 along with their single-objective variants (marked with ∗) in Settings 1–3

	Setting 1			Setting 2			Setting 3		
Alg.	Obj. 1	Obj. 2	Per.	Obj. 1	Obj. 2	Per.	Obj. 1	Obj. 2	Per.
OM 1	12.0 ± 2.1	333 ± 56	97%	321 ± 71	314 ± 61	94%	11.0 ± 2.0	323 ± 60	97%
OM 1*	334 ±73		48%	334 ± 73		48%	334 ± 73		48%
NM 1	1310 ±580	1300 ± 600	74%	1230 ± 660	1180 ± 660	76%	12.4 ± 7.5	1230 ± 640	88%
NM 2	4330 ±3600	2480 ± 2800	32%	3100 ± 2800	3050 ± 2600	38%	14.8 ± 13	4590 ± 3100	54%
NM 3	244 ±140	226 ± 150	95%	249 ± 140	245 ± 160	95%	9.72 ± 5.0	270 ± 120	97%
NM 1*	928 ±780		42%	928 ± 780		42%	928 ± 780		42%
PF 1	764 ±210	723 ± 1.1p	85%	806 ± 240	723 ± 1.1p	85%	679 ± 77	723 ± 1.1p	86%
PF 2	9820 ±4.5	52.8 ± 14f	1.3%	5000 ± 860	94.6 ± 24	50%	52.8 ± 14f	105 ± 32	98%

p denotes $\times 10^{-12}$ and f denotes $\times 10^{-15}$

Means are rounded to three most significant digits, standard deviations are rounded to two most significant digits

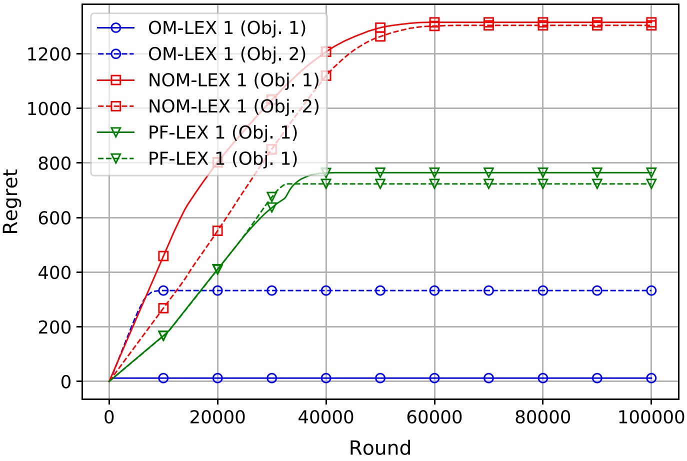

Fig. 1 Regrets of OM-LEX 1, NOM-LEX 1, and PF-LEX 1 in Setting 1

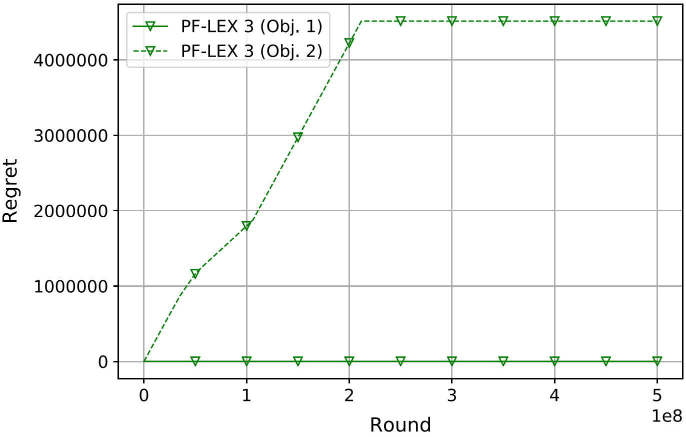

Fig. 2 Regret of PF-LEX 3 in Setting 1

to the time horizon of the problem. In the remaining rounds, it plays only the optimistic near-lexicographic optimal arm in objective 2 ($\hat{a}^2_*(t)$). For this case, ϵ matches with the minimum suboptimality gap. Thus, although PF-LEX always chooses $\hat{a}^2_*(t)$, because $\hat{\mathcal{A}}^1_*(t) = \mathcal{A}^1_*$, it learns to play optimally. As a remark, we note that PF-LEX could incur high regret in objective 1 (see PF-LEX 2 in Table 3) if the minimum suboptimality gap were smaller than ϵ.

Next, for Setting 1, we consider PF-LEX 3 that has parameters $\epsilon = \delta = T^{-1/3}$ as given in Table 2 (that match with the optimal choice for ϵ given in Theorem 1), run simulations for $T = 5 \times 10^8$, and report the average regret of the learner over 5 runs (Fig. 2). This result illustrates the identifiability problem introduced earlier that makes learning lexicographic optimal arms particularly challenging. We see that PF-LEX rules out

Table 4 Prior knowledge and parameters of the algorithms for the two additional settings

Algorithm	Prior knowledge and parameters
OM-LEX 2	$\mu_*^1 = \mu_*^2 = \mu_*^3 = 0.50$
NOM-LEX 4	$\eta_1 = \eta_2 = \eta_3 = 0.45$
NOM-LEX 5	$\eta_1 = \eta_3 = 0.45, \eta_2 = -10^6$

Table 5 Priority-based and priority-free regrets of OM-LEX 2 and NOM-LEX 4–5 in Settings 4–5

	Setting 4			Setting 5		
Algorithms	Obj. 1	Obj. 2	Obj. 3	Obj. 1	Obj. 2	Obj. 3
OM-LEX 2						
pr.-based	2000 ± 100	821 ± 75	367 ± 59	1010 ± 82		373 ± 72
pr.-free	1990 ± 120	1440 ± 110	1290 ± 110	1040 ± 81	−350 ± 17	−350 ± 17
NOM-LEX 4						
pr.-based	6730 ± 2000	2200 ± 810	649 ± 330	7200 ± 2100		1130 ± 440
pr.-free	7000 ± 1500	−4560 ± 3100	−8250 ± 3200	7450 ± 1700	−14,500 ± 3200	−6190 ±1600
NOM-LEX 5						
pr.-based				6570 ± 2700		1060 ± 530
pr.-free				6660 ± 2400	−12,700 ± 5300	−6090 ±2700

Note that the prior information of NOM-LEX 5 is not valid for Setting 4

arm 3 as a potential lexicographic optimal arm and stops incurring regret in objective 1 very early on. However, since it is not possible to be confident in that both arm 1 and arm 2 have equal expected rewards in objective 1, the algorithm still keeps exploring them uniformly until around round 2×10^8. During this exploration stage, it incurs linear regret in objective 2. Once both arms are deemed to be optimal in objective 1, PF-LEX starts exploiting the optimistic near-lexicographic optimal arm in objective 2, after which the increase of the regret in objective 2 drops drastically.

For the two additional settings with $D = 3$, we consider Settings 4 and 5. In Setting 4, there are 43 arms with expected reward vectors in $\{0.90, 0.50, 0.40, 0.10\}^3$ such that each arm has a unique expected reward vector, where we eliminated arms that lexicographically dominate (0.50, 0.50, 0.50) so that it is the only lexicographic optimal arm and $\Delta^1_{\min} = \Delta^2_{\min} = \Delta^3_{\min} = 0.10$. Setting 4 features a large variety of arms with combinations of expected rewards that are much higher than, equal to, slightly lower than, and much lower than the lexicographic optimal expected rewards in all objectives. In Setting 5, there are 19 arms, where we eliminated arms in $\mathcal{S}^2_*$ from Setting 4 so that $\mathcal{S}^2_* = \emptyset$ and $\Delta^2_{\min} = \infty$ while $\Delta^1_{\min} = \Delta^3_{\min} = 0.10$. Table 4 shows OMG-LEX and NOM-LEX with different prior knowledge and parameters than the ones considered so far.

We run simulations with $T = 10^5$ and average the regret of the learners over 100 individual runs. Different from the previous experiments, we provide results for the priority-free regret as well.

Table 6 Prior knowledge and parameters of the algorithms for the Sat-MAB

Algorithm	Prior knowledge and parameters
NOM-LEX 6	$\eta_1 = 2.5$
NOM-LEX 7	$\eta_1 = \eta_2 = \eta_3 = 2.5$
Satisficing-In-Mean-Rewards UCL	$\boldsymbol{\mu}_0 = \mathbf{0}$, $\Sigma_0 = \lim_{\sigma_0^2 \to \infty} \sigma_0^2 I$ $(\Lambda_0 = 0)$, $K = 1$

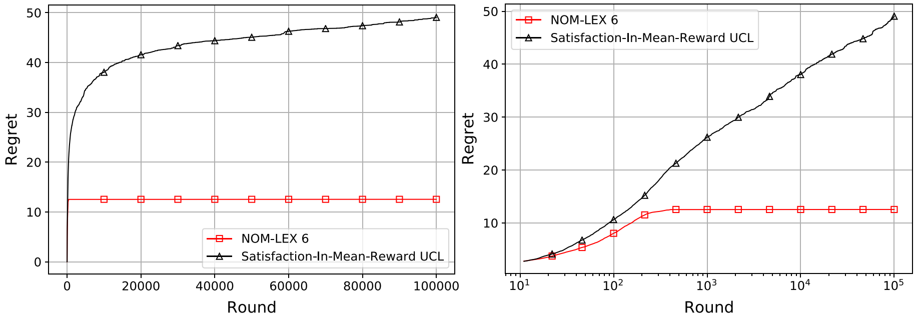

Fig. 3 Regrets of NOM-LEX 6 and Satisficing-In-Mean-Rewards UCL in Setting 6

Table 5 shows the priority-based and priority-free regrets of OM-LEX 2, NOM-LEX 4 and 5 in Settings 4 and 5 at $T = 10^5$. Since NOM-LEX considers arms with very high expected rewards compared to the near optimal expected reward as potential optimal arms, it tends to incur a lot more negative regret in priority-free settings as opposed to OM-LEX, which only looks for arms with expected rewards that are very close to the lexicographic optimal expected rewards.

In Setting 5, note that any $\delta_2 > 0$ would guarantee a bounded regret for NOM-LEX (since $\Delta^2_{\min} = \infty$). Moreover, δ_2 appears in none of our regret bounds in Theorems 5 and 6 for Setting 5. However, our numerical experiments show that it still affects the regret. This is because knowing a larger δ_2 better captures the information $\Delta^2_{\min} = \infty$ and results in having smaller regret in objective 1.

5.2 Experiments for the Sat-MAB

In this section, we demonstrate our results for the Sat-MAB in two new settings: Setting 6 and Setting 7. Setting 6 has four arms and a single objective. The arms have Gaussian rewards with unit variance and expected rewards $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, and $\mu_4 = 4$. The target threshold is set to be 2.5, meaning arms 3 and 4 are satisficing while the other arms are not. Note that Setting 6 is identical to the setting considered in Reverdy et al. (2017) for the satisficing-in-mean-rewards problem (Problem 2 in Reverdy et al., 2017). Setting 7 has 64 arms and three objectives. The arms again have Gaussian rewards with unit variance

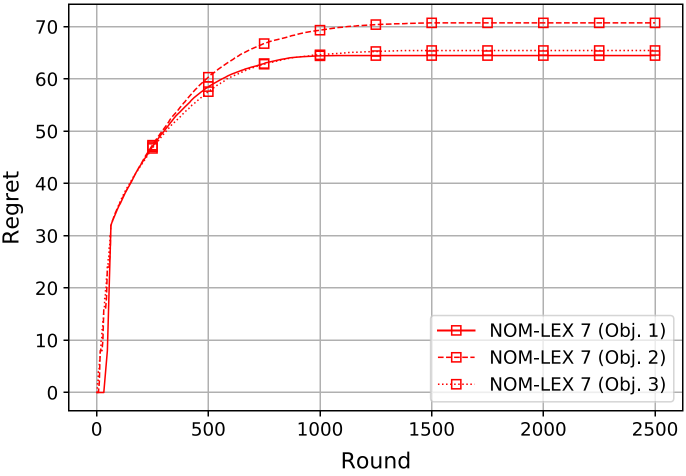

Fig. 4 Regret of NOM-LEX 7 in Setting 7

and unique expected reward vectors in $\{1, 2, 3, 4\}^3$. The target threshold for all objectives is set to be 2.5, meaning there are exactly 8 satisficing arms.

For all cases, we average the regret of learners over 100 individual runs. We consider NOM-LEX with appropriate parameters that are summarized in Table 6. For Setting 6, we also consider Satisficing-In-Mean-Rewards UCL, which is the algorithm proposed in Reverdy et al. (2017). We use the same parameters for UCL as Reverdy et al. (2017), which are also summarized in Table 6.

Figure 3 shows the regret of NOM-LEX 6 and Satisficing-In-Mean-Rewards UCL after $T = 10^5$ rounds. Consistent with the proven regret bounds, NOM-LEX 6 achieves bounded regret while the regret of Satisficing-In-Mean-Rewards UCL grows logarithmically. Figure 4 shows the regret of NOM-LEX 7 after $T = 2500$ rounds. NOM-LEX is not only a better algorithm than Satisficing-In-Mean-Rewards UCL in single-objective settings but also capable of learning in multiobjective settings.

6 Conclusion

We proposed two new multi-objective MAB problems: the Lex-MAB and the Sat-MAB. For the Lex-MAB, we showed that without prior information an almost optimal $\tilde{O}(T^{2/3})$ gap-free regret can be achieved and with prior information the regret is uniformly bounded in time. We also proved that uniformly bounded regret can be achieved for the Sat-MAB as well. The case where there is prior information only for a subset of the objectives is worth investigating in the future.

Table 7 List of system notations

Notation	Definition	Description
A		Number of arms
$\mathcal{A}$	$[A]$	Set of all arms
$a(t)$		Selected arm in round t
D		Number of objectives
$\mathcal{D}$	$[D]$	Set of all objectives
μ_a^i		Expected reward of arm a in objective i
$\boldsymbol{\mu}_a$	$(\mu_a^1, \dots, \mu_a^D)$	Expected reward vector of arm a
$\kappa^i(t)$		Noise in objective i in round t
$r^i(t)$	$\mu_{a(t)}^i + \kappa^i(t)$	Reward of the selected arm in round t
$\succ_{\text{lex},i}$	$\boldsymbol{\mu} \succ_{\text{lex},i} \boldsymbol{\mu'} \iff \mu^j > \mu'^j,\ j = \min\{k \leq i : \mu^k \neq \mu'^k\}$	Symbol for lexicographic dominance in the first i objectives
$\mathcal{A}_*^i$	$\{a : \mathcal{A} : \boldsymbol{\mu}_{a'} \not\succ_{\text{lex},i} \boldsymbol{\mu}_a, \forall a' \in \mathcal{A}\}$	Set of lexicographic optimal arms in the first i objectives
$\mathcal{A}_*$	$\mathcal{A}_*^D$	Set of lexicographic optimal arms in all objectives
$*$	$* : * \in \mathcal{A}_*$	A lexicographic optimal arm
Δ_a^i	$\mu_*^i - \mu_a^i$	Gap of arm a in objective i
∇_a^i	$\lvert \mu_*^i - \mu_a^i \rvert$	Absolute gap of arm a in objective i
$\mathcal{S}_*^i$	$\mathcal{A}_*^{i-1} - \mathcal{A}_*^i$	Set of arms that are lexicographic optimal in the first $i - 1$ objectives but not lexicographic optimal in the first i objectives
$\mathcal{S}^i$	$\{a : \Delta_a^i > 0\}$	Set of suboptimal arms in objective i.
$\Delta_{\max}^i$	$\max_{a \in \mathcal{A}} \Delta_a^i$	Maximum suboptimality gap in objective i
$\nabla_a^{\max}$	$\max_{i \in \mathcal{D}} \nabla_a^i$	Maximum absolute gap of arm a
$\mathbf{Reg}_{pb}(T)$	$(\text{Reg}_{pb}^1(T), \dots, \text{Reg}_{pb}^D(T))$	Lexicographic priority-based regret
$\mathbf{Reg}_{pf}(T)$	$(\text{Reg}_{pf}^1(T), \dots, \text{Reg}_{pf}^D(T))$	Lexicographic priority-free regret
$\text{Reg}_{pb}^i(T)$	$\sum_{t=1}^T \Delta_{a(t)}^i \mathbb{1}\{a(t) \in \mathcal{S}_*^i\}$	Priority-based regret in objective i
$\text{Reg}_{pf}^i(T)$	$\sum_{t=1}^T \Delta_{a(t)}^i$	Priority-free regret in objective i

Table 7 (continued)

Notation	Definition	Description
$\varDelta^i_{\min}$	$\min_{a \in \mathcal{S}^i_*} \varDelta^i_a$ for $\mathbf{Reg}_{pb}(T)$, $\min_{a \in \mathcal{S}^i} \varDelta^i_a$ for $\mathbf{Reg}_{pf}(T)$	Minimum suboptimality gap in objective i

Table 8 List of notations for Case 1

Notation	Definition	Description
ϵ		Suboptimality that is aimed to be tolerated
δ		Confidence term
$N_a(t)$	$\sum_{t'=1}^{t-1} \mathbb{I}\{a(t') = a\}$	Number of times arm a was selected by the beginning of round t
$\hat{\mu}_a^i(t)$	$\sum_{t'=1}^{t-1} r^i(t')/N_a(t)$	Sample mean of the rewards of arm a in objective i at the beginning of round t
$c_a(t)$	See Sect. 4.1	Half of the length of the confidence interval of arm a at the beginning of round t
$u_a^i(t)$	$\hat{\mu}_a^i(t) + c_a(t)$	Upper confidence bound of arm a in objective i at the beginning of round t
$l_a^i(t)$	$\hat{\mu}_a^i(t) - c_a(t)$	Lower confidence bound of arm a in objective i at the beginning of round t
$C_{i,t}$	See Sect. 4.1	Chains to in objective i in round t
$\hat{a}_*^i(t)$	$\text{argmax}_{a \in \hat{\mathcal{A}}_*^{i-1}(t)} u_a^i(t)$	Optimistic near-lexicographic optimal arm in objective i at the beginning of round t
$\hat{\mathcal{A}}_*^i(t)$	$\{a \in \mathcal{A} : a\, C_{i,t}\, \hat{a}_*^i(t)\}$	Set of estimated lexicographic optimal arms in the first i objectives at the beginning of round t

Table 9 List of notations for Case 2

Notation	Definition	Description
$N_a(t)$	$\sum_{t'=1}^{t-1} \mathbb{I}\{a(t') = a\}$	Number of times arm a was selected by the beginning of round t
$\hat{\mu}_a^i(t)$	$\sum_{t'=1}^{t-1} r^i(t')/N_a(t)$	Sample mean of the rewards of arm a in objective i at the beginning of round t
$\hat{\mathcal{A}}_*(t)$	$\{a \in \mathcal{A} : \forall i \in \mathcal{D}, \lvert\hat{\mu}_a^i(t)\rvert < \sqrt{4 \log N_a(t)/N_a(t)}\}$	Set of estimated lexicographic optimal arms at the beginning of round t
$\dagger(a)$	$\operatorname{argmax}_{i \in \mathcal{D}} \nabla_a^i$	Objective for which $\nabla_a^{\dagger(a)} = \nabla_a^{\max}$
$(\cdot)_a^\dagger$	$(\cdot)_a^{\dagger(a)}$	

Table 10 List of notations for Case 3

Notation	Definition	Description
η_i	$\eta_i : \mu_*^i - \Delta_{\min}^i < \eta_i < \mu_*^i$	Near-lexicographic optimal expected reward in objective i
δ_i	$\mu_*^i - \eta_i$	Gap between near-lexicographic optimal expected reward and the lexicographic optimal expected reward in objective i
$N_a(t)$	$\sum_{t'=1}^{t-1} \mathbb{I}\{a(t') = a\}$	Number of times arm a was selected by the beginning of round t
$\hat{\mu}_a^i(t)$	$\sum_{t'=1}^{t-1} r^i(t')/N_a(t)$	Sample mean of the rewards of arm a in objective i at the beginning of round t
$\hat{\mathcal{A}}_*(t)$	$\{a \in \mathcal{A} : \forall i \in \mathcal{D}, \hat{\mu}_a^i(t) > -\sqrt{4 \log N_a(t)/N_a(t)}\}$	Set of estimated lexicographic optimal arms at the beginning of round t
$\dagger(a)$	$\operatorname{argmax}_{i \in \mathcal{D}}(\Delta_a^i - \delta_i)$	
$(\cdot)_a^\dagger$	$(\cdot)_a^{\dagger(a)}$	

Appendix: Tables of Notation

General notation is listed in Table 7. Notations specific to each case covered in Section 4 are listed in Tables 8, 9 and 10 respectively.

Acknowledgements The work of Cem Tekin was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 215E342 and in part by the BAGEP Award of the Science Academy.

References

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in neural information processing systems* (pp. 2312–2320).

Antos, A., Grover, V., & Szepesvári, C. (2010). Active learning in heteroscedastic noise. *Theoretical Computer Science*, *411*(29–30), 2712–2728.

Bubeck, S., & Liu, C. (2013). Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in neural information processing systems* (pp 638–646)

Bubeck, S., Perchet, V., & Rigollet, P. (2013). Bounded regret in stochastic multi-armed bandits. In *Proceedings of the conference on learning theory* (pp. 122–134).

Cully, A., Clune, J., Tarapode, D., & Mouret, J. (2015). Robots that can adapt like animals. *Nature*, *521*(7553), 503–507.

Drugan, M., & Nowe, A. (2013). Designing multi-objective multi-armed Bandtis algorithms: A study. In *Proceedings of the 2013 international joint conference on neural networks* (pp. 1–8).

Ehrgott, M. (2005). *Multicriteria optimization* (Vol. 491). Springer.

Fishburn, P. (1974). Exceptional paper–lexicographic orders, utilities and decision rules: A survey. *Management Science*, *20*(11), 1442–1472.

Gai, Y., Krishnamachari, B., & Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, *20*(5), 1466–1478.

Garivier, A., Ménard, P., & Stoltz, G. (2018). Explore first, exploit next, the true shape of regret in bandit problems. *Mathematics of Operations Research*.

Jee, K., McShan, D., & Fraass, B. (2007). Lexicographic ordering: intuitive multicriteria optimization for imrt. *Physics in Medicine & Biology, 52*(7).

Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2016). Fairness in learning: classic and contextual bandits. In *Advances in neural information processing systems* (pp. 325–333).

Konstan, J., McNee, S., Ziegler, C., Torres, R., Kapoor, N., & Riedl, J. (2006). Lessons on applying automated recommender systems to information-seeking tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *6*, 1630–1633.

Lai, T., & Robbins, H. (1984). Optimal sequential sampling from two populations. *Proceedings of the National Academy of Sciences of the Unites States of America*, *81*(4), 1284–1286.

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

Lattimore, T., & Munos, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in neural informations processing systems* (pp. 550–558).

Lattimore, T., & Szepesvári, C. (2019). *Bandit algorithms*. Cambridge University Press. Preprint.

Locatelli, A., Gutzeit, M., & Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd international conference on machine learning* (pp. 1690–1698).

Mersereau, A., Rusmevichientong, P., & Tsitsiklis, J. (2009). A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, *54*(12), 2787–2802.

Reverdy, P., Srivastava, V., & Leonard, N. (2017). Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, *62*(8), 3788–3803.

Shah-Mansouri, V., Mohsenian-Rad, A., & Wong, V. (2009). Lexicographically optimal routing for wireless sensor networks with multiple sinks. *IEEE Transactions on Vehicular Technology*, *58*(3), 1490–1500.

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research*, *15*(1), 1673–1681.

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

Turgay, E., Öner, D., & Tekin, C. (2018). Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st international conference on artificial intelligence and statistics* (pp. 1673–1681).

Vakili, S., & Zhao, Q. (2013). Achieving complete learning in multi-armed bandit problems. In *Proceedings of the 2013 asilomar conference on signals, systems and computers* (pp. 1778–1782).

Zhou, T., Kuscsik, Z., Liu, J., Medo, M., Wakeling, J., & Zhang, Y. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, *107*, 4511–4515.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.