Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1065 – 1080

doi:10.3906/elk-1806-221

The biobjective multiarmed bandit: learning approximate lexicographic optimal allocations

Cem TEKİN*
Department of Electrical and Electronics Engineering, Faculty of Engineering, Bilkent University, Ankara, Turkey

Received: 29.06.2018 • **Accepted/Published Online:** 10.12.2018 • **Final Version:** 22.03.2019

Abstract: We consider a biobjective sequential decision-making problem where an allocation (arm) is called ϵ lexicographic optimal if its expected reward in the first objective is at most ϵ smaller than the highest expected reward, and its expected reward in the second objective is at least the expected reward of a lexicographic optimal arm. The goal of the learner is to select arms that are ϵ lexicographic optimal as much as possible without knowing the arm reward distributions beforehand. For this problem, we first show that the learner's goal is equivalent to minimizing the ϵ lexicographic regret, and then, propose a learning algorithm whose ϵ lexicographic gap-dependent regret is bounded and gap-independent regret is sublinear in the number of rounds with high probability. Then, we apply the proposed model and algorithm for dynamic rate and channel selection in a cognitive radio network with imperfect channel sensing. Our results show that the proposed algorithm is able to learn the approximate lexicographic optimal rate–channel pair that simultaneously minimizes the primary user interference and maximizes the secondary user throughput.

Key words: Multiarmed bandit, biobjective learning, lexicographic optimality, dynamic rate and channel selection, cognitive radio networks

1. Introduction

The multiarmed bandit (MAB) is used to model real-world applications in which the decision maker repeatedly interacts with its unknown environment in order to maximize its long-term reward [1, 2]. The decision maker can be a recommender system recommending items to its users [3], a secondary user performing opportunistic spectrum access in a cognitive radio network [4], or an agent that chooses a routing path between the source and the destination in a network [5].

A plethora of prior works on the MAB focused on designing learning algorithms that optimize the total scalar reward. These include the celebrated upper confidence bound (UCB) policies [1, 6] and posterior sampling [2, 7]. On the other hand, in many real-world applications of the MAB, the environment produces vector-valued rewards, where each component of the reward vector corresponds to a different goal. For instance, in a cognitive radio network, the goal of the secondary user (SU) is to maximize its throughput while minimizing the interference to the primary user (PU). In this paper, we introduce the biobjective MAB to tackle this type of sequential decision-making problems. In the biobjective MAB, the learner receives, at each round, random rewards from two objectives. These objectives are lexicographically ordered in the sense that the learner values the first objective more than the second objective.

The learner aims at selecting approximate lexicographic optimal allocations (arms), which yield an ϵ

*Correspondence: cemtekin@ee.bilkent.edu.tr

optimal expected reward in the first objective and an expected reward that is at least the expected reward of a lexicographic optimal arm in the second objective. This notion of optimality allows the learner to accumulate a high reward in the second objective by incurring a small loss in the first objective. In order to quantify the loss of the learner due to not knowing the ϵ lexicographic optimal arms beforehand, we introduce the notion of ϵ lexicographic regret, and propose a learning algorithm whose ϵ lexicographic gap-dependent regret is $O(1)$ and gap-independent regret is $\tilde{O}(\sqrt{T})$ with high probability. Then, we cast the dynamic rate and channel selection problem in a cognitive radio network with imperfect channel sensing as a biobjective MAB where the first objective is related to PU interference and the second objective is related to SU throughput.

To sum up, in this work, we propose a new MAB called the biobjective MAB, study the notion of approximate lexicographic optimality, propose a learning algorithm and bound its regret, and investigate a multirate multichannel communication application of the biobjective MAB. The algorithm we propose is fundamentally different from the algorithms designed to learn in the MAB with scalar reward and uses confidence intervals, in addition to the UCBs, in order to learn the optimal arms based on lexicographic ordered objectives. This also makes the regret analysis substantially different from the prior work, since bounding the regret requires considering the confidence intervals for both objectives.

2. Related work

In the classical MAB, first studied in [2], at each round, after selecting an arm, the learner receives a random reward that comes from an unknown distribution that depends on the selected arm. An asymptotically optimal adaptive allocation rule with $O(\log T)$ regret is proposed in [1] for the classical MAB with independent arms. Later, finite time $O(\log T)$ regret bounds are derived in [6]. It is also shown in [1] that when the arms are independent, the best possible regret is $O(\log T)$. Numerous interesting extensions of the classical MAB are proposed later on, including the combinatorial MAB [8] and the unimodal MAB [9, 10].

For instance, [8] proposes the combinatorial MAB in which the learner selects at each round a super arm that is composed of multiple arms, observes the outcomes of the selected arms, and receives a linear combination of the rewards of the selected arms. The combinatorial bandit is used in [11] to learn the optimal allocations in a multiuser multichannel communication system. Due to obtaining observations from each selected arm, this problem is also called the combinatorial semibandit [12]. Reward functions that are nonlinear in the expected outcomes of arms are considered in [13] and [14].

The variant of the classical MAB we consider in this paper is the multiobjective MAB. Unlike the classical MAB, where the reward is scalar, the reward is vector valued in the multiobjective MAB. This results in various notions of optimality, each of which require a different learning algorithm. For instance, Pareto optimality is considered in [15], [16], and [17]. Essentially, an arm is called Pareto optimal, if switching to any arm that is better in terms of the expected reward in at least one objective will result in a reduction in the expected reward in at least one other objective. It is shown that the Pareto regret, i.e. the loss due to not selecting arms from the Pareto front, is $O(\log T)$. As an extension, contextual multiobjective MAB with similarity information is considered in [18]. In this work, the authors propose a multiobjective learning algorithm that uses the contextual zooming idea [19], and prove that the Pareto regret is $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ where d_p is the Pareto zooming dimension of the similarity space.

Another important notion of optimality in the multiobjective setting is lexicographic optimality [20]. Unlike Pareto optimality in lexicographic optimality, the order of the objectives matter. In this case, the

learner prefers obtaining higher reward in any objective i to obtaining higher reward in any other objective j such that $i < j$. Lexicographic optimality is first studied in a contextual MAB [21, 22], and it is shown that the lexicographic regret is $\tilde{O}(T^{(2+d)/(3+d)})$, where d is the dimension of the context set. In addition to the MAB, the notions of Pareto optimality and lexicographic optimality are also considered in the more general reinforcement learning framework [23, 24].

Compared to all the works mentioned above, in this paper, we propose the biobjective MAB with approximate lexicographic optimality as the performance metric for the first time. As opposed to lexicographic optimality, we analyze the learner's performance when it can tolerate $\epsilon > 0$ suboptimality in the first objective. This way, the learner seeks to identify and select ϵ optimal arms in the first objective, which might result in significant improvement in the reward it obtains in the second objective. We prove two high probability bounds on the ϵ lexicographic regret: $O(1)$ gap-dependent regret bound and $\tilde{O}(\sqrt{T})$ gap-independent regret bound. These bounds are much sharper than the $\tilde{O}(T^{(2+d)/(3+d)})$ regret bound for the multiobjective contextual MAB, since the existence of contexts makes learning of lexicographic optimal allocations more difficult.

3. Problem formulation

In this section, we explain the system model, and define approximate lexicographic optimality and the regret. Our notation is presented in Table 1.

Table 1. Notation

Notation for problem description			
$\mathcal{A}$	Set of arms	$a(t)$	Arm selected in round t
μ_a^i	Expected reward of arm a in obj. i	$r^i(t)$	Random reward in obj. i
$\boldsymbol{\mu}_a$	Expected reward vector of arm a	$\kappa^i(t)$	Noise in obj. i
$\mathcal{A}_*^1$	Set of arms with the highest expected	μ_*^1	The highest expected reward
	reward in obj. 1		in obj. 1
$\mathcal{A}_*^2$	Set of lexicographic optimal arms	μ_*^2	Expected reward of an arm in $\mathcal{A}_*^2$ in obj. 2
$\Delta_{a,\epsilon}^1$	Suboptimality gap of arm a in obj. 1	Δ_a^2	Suboptimality gap of arm a in obj. 2
$\mathcal{S}_i$	Set of suboptimal arms in obj. i	$\mathrm{Reg}_\epsilon^1(T)$	Regret in obj. 1
$\mathrm{Reg}^2(T)$	Regret in obj. 2	$\mathbf{Reg}_\epsilon(T)$	ϵ lexicographic regret
Notation for the learning algorithm (ALEX)			
$N_a(t)$	Number of times arm a was selected	$\hat{\mu}_a^i(t)$	Sample mean estimate of μ_a^i
	prior to round t		in round t
$u_a^i(t)$	Upper confidence bound (UCB) for	$l_a^i(t)$	Lower confidence bound (LCB) for
	the expected reward in obj. i		the expected reward in obj. i
$c_a(t)$	Uncertainty term	$\hat{a}_*^1(t)$	Arm with the highest UCB in obj. 1
$\hat{\mathcal{A}}_*^1(t)$	Set of candidate optimal arms		

3.1. System model

We consider decision epochs (rounds) indexed by $t \in \{1, 2, \ldots\}$. At each round t, the learner first selects an arm $a(t)$ from the finite arm set $\mathcal{A}$, and then, observes a random reward for each objective $i \in \{1, 2\}$,

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 μ^i_a, $a \in \mathcal{A}$ beforehand, and the noise process $(\kappa^1(t), \kappa^2(t))$ is assumed to be independent over rounds and conditionally 1-sub-Gaussian, i.e. $\forall \lambda \in \mathbb{R}$ $\mathrm{E}[e^{\lambda \kappa^i(t)} | a(t)] \leq \exp(\lambda^2/2)$. This 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. We use $\boldsymbol{\mu}_a := (\mu^1_a, \mu^2_a)$ to denote the expected reward vector of arm a.

3.2. Approximate lexicographic optimality

Let $\mathcal{A}^1_* := \arg\max_{a \in \mathcal{A}} \mu^1_a$ denote the set of arms with the highest expected reward and $\mu^1_* := \max_{a \in \mathcal{A}} \mu^1_a$ denote the highest expected reward in objective 1. The set of lexicographic optimal arms is defined as $\mathcal{A}^2_* := \arg\max_{a \in \mathcal{A}^1_*} \mu^2_a$. The expected reward of a lexicographic optimal arm in objective 2 is defined as $\mu^2_* := \max_{a \in \mathcal{A}^1_*} \mu^2_a$. Moreover, when referring to a lexicographic optimal arm we use a_*. For a given $\epsilon > 0$, arm a is called ϵ (approximate) lexicographic optimal if it satisfies the following condition: $\mu^1_a \geq \mu^1_* - \epsilon$ and $\mu^2_a \geq \mu^2_*$. We define the suboptimality gap of arm a in objective 1 as $\Delta^1_{a,\epsilon} := [\mu^1_* - \mu^1_a - \epsilon]_+$ and in objective 2 as $\Delta^2_a := [\mu^2_* - \mu^2_a]_+$, where $[\mu]_+ = \max\{0, \mu\}$. Based on this, the set of suboptimal arms in objectives 1 and 2 are defined as $\mathcal{S}_1 := \{a \in \mathcal{A} : \Delta^1_{a,\epsilon} > 0\}$ and $\mathcal{S}_2 := \{a \in \mathcal{A} : \Delta^2_a > 0\}$ respectively. Cardinalities of these sets are represented by using $|\cdot|$. For instance, $|\mathcal{S}_1|$ represents the cardinality of $\mathcal{S}_1$.

In many learning applications, it is intuitive to consider approximate lexicographic optimality instead of lexicographic optimality. For instance, when there are many near-optimal arms in objective 1, an arm which is slightly worse than the best arm in objective 1 can have a much higher expected reward in objective 2 than the best arm in objective 1. Such a case is considered in Section 6.

3.3. Regret definition

Since the learner does not know the expected arm rewards beforehand, we compare it with an oracle, which knows the expected rewards of the arms and chooses an ϵ lexicographic optimal arm in each round. The loss of the learner with respect to this oracle is measured by the ϵ lexicographic (pseudo) regret (referred to as the regret hereafter), and is given as the tuple $\mathbf{Reg}_\epsilon(T) := (\mathrm{Reg}^1_\epsilon(T), \mathrm{Reg}^2(T))$, where

$$\mathrm{Reg}^1_\epsilon(T) := \sum_{t=1}^{T} \Delta^1_{a(t),\epsilon} \quad \text{and} \quad \mathrm{Reg}^2(T) := \sum_{t=1}^{T} \Delta^2_{a(t)}. \tag{1}$$

Using the multidimensional regret notion defined above, we say that $\mathbf{Reg}_\epsilon(T)$ is $O(\max\{f_1(T), f_2(T)\})$ when $\mathrm{Reg}^1_\epsilon(T) = O(f_1(T))$ and $\mathrm{Reg}^2(T) = O(f_2(T))$. In Section 4, we propose a learning algorithm with a gap-dependent regret of $O(1)$ with high probability and $O(\log T)$ in expectation, and a gap-independent regret of $\tilde{O}(\sqrt{T})$ both with high probability and in expectation. The difference between the gap-dependent and the gap-independent regrets is that the former depends on problem-specific parameters such as the minimum suboptimality gap, while the latter does not have any dependence on such parameters (i.e. it holds for the worst-case selection of problem-specific parameters).

4. The learning algorithm

Our algorithm is named Approximate Lexicographic Exploration and Exploitation (ALEX) and its pseudocode is given in Algorithm 1. ALEX takes as input $\epsilon > 0$ and for each arm a it keeps a counter $N_a(t)$, which counts the number of times arm a was selected prior to round t, and the sample mean estimate of the rewards from the selections of arm a prior to round t for objectives 1 and 2, denoted by $\hat{\mu}_a^1(t)$ and $\hat{\mu}_a^2(t)$ respectively.

Arm selection of ALEX in round t depends on the confidence intervals in the first objective. The upper confidence bound (UCB) and the lower confidence bound (LCB) of arm a in objective i are given as $u_a^i(t) := \hat{\mu}_a^i(t) + c_a(t)$ and $l_a^i(t) := \hat{\mu}_a^i(t) - c_a(t)$ respectively. Here,

$$c_a(t) = \sqrt{\frac{1+N_a(t)}{N_a^2(t)}\left(1+2\log\left(\frac{2|\mathcal{A}|(1+N_a(t))^{1/2}}{\delta}\right)\right)} \tag{2}$$

represents the uncertainty in arm a's reward, and δ is called the confidence term, which is given as input to ALEX. As expected, the uncertainty decreases as arm a gets selected. As we will show in Section 5, μ_a^i is in the confidence interval $[l_a^i(t), u_a^i(t)]$ with high probability for both objectives in all rounds. Let $\hat{a}_*^1(t) := \arg\max_{a\in\mathcal{A}} u_a^1(t)$ denote the arm with the highest UCB in objective 1. The confidence bounds imply that an arm a for which $u_a^1(t) < l_{\hat{a}_*^1(t)}^1(t) - \epsilon/3$ is suboptimal in the first objective with high probability. Thus, the set of candidate optimal arms in round t is defined as

$$\hat{\mathcal{A}}_*^1(t) := \left\{a \in \mathcal{A} : u_a^1(t) \geq l_{\hat{a}_*^1(t)}^1(t) - \epsilon/3\right\}. \tag{3}$$

When the uncertainty about arm $\hat{a}_*^1(t)$ is high, i.e. $c_{\hat{a}_*^1(t)}(t) > \epsilon/3$, ALEX selects arm $a(t) = \hat{a}_*^1(t)$ to reduce its uncertainty. However, since this selection does not take into account the rewards obtained in objective 2, it does not ensure selection of ϵ lexicographic optimal arms. On the other hand, when the uncertainty about arm $\hat{a}_*^1(t)$ is low, i.e. $c_{\hat{a}_*^1(t)}(t) \leq \epsilon/3$, ALEX selects the arm in $\hat{\mathcal{A}}_*^1(t)$ with the highest UCB in objective 2, i.e. $a(t) = \arg\max_{a\in\hat{\mathcal{A}}_*^1(t)} u_a^2(t)$. This ensures that an ϵ lexicographic optimal arm is selected with high probability.

After ALEX selects arm $a(t)$, it observes the random reward vector $\boldsymbol{r}(t) = (r^1(t), r^2(t))$ of arm $a(t)$, and updates the sample mean estimates of the rewards in objectives 1 and 2 and the counter of $a(t)$. This procedure is repeated in the next round.

5. Regret analysis

In this section, we prove $O(1)$ gap-dependent and $\tilde{O}(\sqrt{T})$ gap-independent regret bounds for ALEX in the event that the confidence intervals hold. We also show that the confidence intervals hold with high probability, which allows us to translate the bounds that we derive for regret to the expected regret. The biobjective nature of the problem requires us to analyze the regrets incurred in objectives 1 and 2 separately. Essentially, for the regret in objective 2, we need to deal with two cases: the case where ALEX forces selection of $\hat{a}_*^1(t)$ and the case where ALEX selects an arm from its candidate optimal arm set $\hat{\mathcal{A}}_*^1(t)$.

Throughout our analysis, complement of event $\mathcal{E}$ is denoted by $\mathcal{E}^c$. First, we state a concentration inequality that will be used in the proofs.

Algorithm 1 ALEX

Input: ϵ, δ
Initialize counters: $N_a = 0$, $\forall a \in \mathcal{A}$, $t = 1$
Initialize estimates: $\hat{\mu}_a^1 = \hat{\mu}_a^2 = 0$, $\forall a \in \mathcal{A}$
while $t \geq 1$ **do**
  Compute $u_a^i = \hat{\mu}_a^i + c_a$ and $l_a^i = \hat{\mu}_a^i - c_a$ for $a \in \mathcal{A}$, $i \in \{1,2\}$
  Set $\hat{a}_*^1 = \arg\max_{a \in \mathcal{A}} u_a^1$ (ties are broken randomly)
  if $c_{\hat{a}_*^1} > \epsilon/3$ **then**
    Select arm $a(t) = \hat{a}_*^1$
  else
    Compute candidate optimal arms: $\hat{\mathcal{A}}_*^1 = \{a \in \mathcal{A} : u_a^1 \geq l_{\hat{a}_*^1}^1 - \epsilon/3\}$
    Select arm $a(t) = \arg\max_{a \in \hat{\mathcal{A}}_*^1} u_a^2$ (ties are broken randomly)
  end if
  Observe the random reward vector $\boldsymbol{r}(t) = (r^1(t), r^2(t))$
  Update estimates: $\hat{\mu}_{a(t)}^i \leftarrow (\hat{\mu}_{a(t)}^i N_{a(t)} + r^i(t))/(N_{a(t)} + 1)$, $i \in \{1,2\}$
  Update counters: $N_{a(t)} \leftarrow N_{a(t)} + 1$
  $t \leftarrow t + 1$
end while

Lemma 1 *(Lemma 6 in [25]) Consider an arm a for which the rewards of objective i are generated by a process $\{R_a^i(t)\}_{t=1}^T$ with $\mu_a^i = E[R_a^i(t)]$, where the noise $R_a^i(t) - \mu_a^i$ is conditionally 1-sub-Gaussian. Let $N_a(T)$ denote the number of times a is selected by the beginning of round T. Let $\hat{\mu}_a(T) = \sum_{t=1}^{T-1} I(a(t) = a) R_a^i(t)/N_a(T)$ for $N_a(T) > 0$ and $\hat{\mu}_a(T) = 0$ for $N_a(T) = 0$. Then, for any $0 < \delta < 2|\mathcal{A}|$ with probability at least $1 - \delta/(2|\mathcal{A}|)$ we have*

$$|\hat{\mu}_a(T) - \mu_a| \leq \sqrt{\frac{1 + N_a(T)}{N_a^2(T)} \left(1 + 2\log\left(\frac{2|\mathcal{A}|(1 + N_a(T))^{1/2}}{\delta}\right)\right)} \quad \forall T \in \mathbb{N}. \tag{4}$$

Next, we define events in which confidence intervals are violated in at least one round. 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 $\mathrm{UC} := \cup_{i \in \{1,2\}} \mathrm{UC}^i$. The following lemma shows that UC occurs with a very little probability.

Lemma 2 $\Pr(UC) \leq \delta$.

Proof This follows from the concentration inequality given in 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 $(\mathrm{UC}_a^i)^c$ holds with probability at least $1 - \delta/(2|\mathcal{A}|)$, and hence, UC_a^i holds with probability at most $\delta/(2|\mathcal{A}|)$. From the union bound it follows that $\Pr(\mathrm{UC}) \leq \delta$. □

The next lemma bounds for event UC^c the difference between the expected reward of the selected arm and the expected reward of a lexicographic optimal arm in objective 1 as a function of ϵ and the length of the confidence interval of the selected arm.

Lemma 3 *When ALEX is run, the following holds for event UC^c: $\mu_*^1 - \mu_{a(t)}^1 \leq u_{a(t)}^1(t) - l_{a(t)}^1(t) + \epsilon$ for all $t \in \{1, \ldots, T\}$.*

Proof For event UC^c, we have

$$\mu^1_* - \mu^1_{a(t)} \leq u^1_{a_*}(t) - l^1_{a(t)}(t) \tag{5}$$

$$\leq u^1_{\hat{a}^1_*(t)}(t) - l^1_{a(t)}(t) \tag{6}$$

$$\leq u^1_{a(t)}(t) - l^1_{a(t)}(t) + \epsilon. \tag{7}$$

Here, Eq. (5) holds since $\mu^1_* \leq u^1_{a_*}(t)$ and $\mu^1_{a(t)} \geq l^1_{a(t)}(t)$ for event UC^c, Eq. (6) holds since $u^1_{\hat{a}^1_*(t)}(t) \geq u^1_{a_*}(t)$ for all t by definition of $\hat{a}^1_*(t)$, and Eq. (7) holds since $u^1_{a(t)}(t) \geq u^1_{\hat{a}^1_*(t)}(t) - \epsilon$ for all t. For the last inequality, observe that when $c_{\hat{a}^1_*(t)}(t) \leq \epsilon/3$, by the arm selection rule of ALEX we have $u^1_{a(t)}(t) \geq l^1_{\hat{a}^1_*(t)}(t) - \epsilon/3 = u^1_{\hat{a}^1_*(t)}(t) - 2c_{\hat{a}^1_*(t)}(t) - \epsilon/3 \geq u^1_{\hat{a}^1_*(t)}(t) - \epsilon$, and when $c_{\hat{a}^1_*(t)}(t) > \epsilon/3$, again by the arm selection rule of ALEX $a(t) = \hat{a}^1_*(t)$, thus we have $u^1_{a(t)}(t) = u^1_{\hat{a}^1_*(t)}(t) \geq u^1_{\hat{a}^1_*(t)}(t) - \epsilon$. □

Let $\mathcal{T} := \{1 \leq t \leq T : c_{\hat{a}^1_*(t)} \leq \epsilon/3\}$ denote the set of rounds in which ALEX selects an arm based on the UCBs in objective 2 (lines 10–11 of Algorithm 1) and $\mathcal{T}^c := \{1, \ldots, T\} - \mathcal{T}$. In the following lemma, the suboptimality gap of the arm selected in round $t \in \mathcal{T}$ in objective 2 is bounded for event UC^c by the length of the confidence interval of the selected arm.

Lemma 4 *When ALEX is run, the following holds for event UC^c: $\mu^2_* - \mu^2_{a(t)} \leq u^2_{a(t)}(t) - l^2_{a(t)}(t)$ for $t \in \mathcal{T}$.*

Proof Consider any lexicographic optimal arm a_*. For event UC^c, we have $u^1_{a_*}(t) \geq \mu^1_* \geq \mu^1_{\hat{a}^1_*(t)} \geq l^1_{\hat{a}^1_*(t)}(t)$, which implies that $a_* \in \hat{\mathcal{A}}^1_*(t)$. Thus, we have

$$\mu^2_* - \mu^2_{a(t)} \leq u^2_{a_*}(t) - l^2_{a(t)}(t) \tag{8}$$

$$\leq u^2_{a(t)}(t) - l^2_{a(t)}(t), \tag{9}$$

where Eq. (8) holds since $\mu^2_* \leq u^2_{a_*}(t)$ and $\mu^2_{a(t)} \geq l^2_{a(t)}(t)$ for event UC^c, and Eq. (9) holds since $u^2_{a(t)}(t) \geq u^2_{a_*}(t)$ by the arm selection rule of ALEX on $t \in \mathcal{T}$. □

We also need to bound the regret in objective 2 for rounds up to round T for which $t \notin \mathcal{T}$. Let $\mathcal{T}^c_a := \{t \in \{1, \ldots, T\} - \mathcal{T} : \hat{a}^1_*(t) = a\}$. Obviously, ALEX does not incur any regret in objective 2 in rounds $t \in \mathcal{T}^c_a$ for $a \in \mathcal{A} - \mathcal{S}_2$, and incurs regret Δ^2_a in objective 2 in rounds $t \in \mathcal{T}^c_a$ for $a \in \mathcal{S}_2$.

Lemma 5 *When ALEX is run, we have*

$$\sum_{t \in \mathcal{T}^c} \Delta^2_{a(t)} \leq \sum_{a \in \mathcal{S}_2} \left(3 + \frac{36}{\epsilon^2} \log \frac{6e^{\frac{1}{2}}|\mathcal{A}|}{\epsilon\delta}\right) \Delta^2_a. \tag{10}$$

Proof The proof follows from bounding the cardinality of $\mathcal{T}^c_a$ for $a \in \mathcal{S}_2$. Note that $t \in \mathcal{T}^c_a$ when $c_a(t) > \epsilon/3$. Similar to the proof of Theorem 7 in [25], this implies that

$$\frac{N_a^2(t)-1}{N_a(t)+1} \leq \frac{N_a^2(t)}{N_a(t)+1} \leq \frac{9}{\epsilon^2}\left(2\log\frac{2e^{\frac{1}{2}}|\mathcal{A}|(1+N_a(t))^{\frac{1}{2}}}{\delta}\right) \tag{11}$$

Then, from Lemma 8 in [26], we obtain $N_a(t) \leq 3+\frac{36}{\epsilon^2}\log\frac{6e^{\frac{1}{2}}|\mathcal{A}|}{\epsilon\delta}$.

□

In the rest of the analysis, we will bound both $\mathrm{Reg}^i(T)$ under the event UC^c and $\mathrm{E}[\mathrm{Reg}^i(T)]$ by using the results of the lemmas above. For the latter, we will use the following decomposition:

$$\mathrm{E}[\mathrm{Reg}^i(T)] = \mathrm{E}[\mathrm{Reg}^i(T)|\mathrm{UC}]\Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i(T)|\mathrm{UC}^c]\Pr(\mathrm{UC}^c) \leq T\Delta^i_{\max}\Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i_p(T)|\mathrm{UC}^c], \tag{12}$$

where $\Delta^1_{\max} = \max_{a\in\mathcal{A}}\Delta^1_{a,\epsilon}$ and $\Delta^2_{\max} = \max_{a\in\mathcal{A}}\Delta^2_a$.

The following theorem gives gap-dependent regret bounds for ALEX.

Theorem 1 *When ALEX is run with $\delta\in(0,1)$ and $\epsilon>0$, the following bounds hold with probability at least $1-\delta$ for all $T>0$:*

$$Reg^1_\epsilon(T) \leq \sum_{a:\Delta^1_{a,\epsilon}>0}\left(3\Delta^1_{a,\epsilon}+\frac{16}{\Delta^1_{a,\epsilon}}\log\left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\Delta^1_{a,\epsilon}\delta}\right)\right), \tag{13}$$

$$Reg^2(T) \leq \sum_{a:\Delta^2_a>0}\left(3\Delta^2_a+\frac{16\Delta^2_a}{(\min\{\Delta^2_a,2\epsilon/3\})^2}\log\left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\min\{\Delta^2_a,2\epsilon/3\}\delta}\right)\right). \tag{14}$$

Moreover, when ALEX is run with $\delta=1/T$, we have the following bounds on the expected regret:

$$E[Reg^1_\epsilon(T)] \leq \sum_{a:\Delta^1_{a,\epsilon}>0}\left(3\Delta^1_{a,\epsilon}+\frac{16}{\Delta^1_{a,\epsilon}}\log\left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|T}{\Delta^1_{a,\epsilon}}\right)\right)+\Delta^1_{\max}, \tag{15}$$

$$E[Reg^2(T)] \leq \sum_{a:\Delta^2_a>0}\left(3\Delta^2_a+\frac{16\Delta^2_a}{(\min\{\Delta^2_a,2\epsilon/3\})^2}\log\left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|T}{\min\{\Delta^2_a,2\epsilon/3\}}\right)\right)+\Delta^2_{\max}. \tag{16}$$

Proof We first bound the regret in objective 1. For event UC^c, if arm a is selected in round t, then we have $c_a(t)\geq\Delta^1_{a,\epsilon}/2$ (by Lemma 3). The rest of the proof is similar to the proof of Theorem 7 of [25]:

$$c_a(t)\geq\frac{\Delta^1_{a,\epsilon}}{2} \Rightarrow N_a(t)\leq 3+\frac{16}{(\Delta^1_{a,\epsilon})^2}\log\left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\Delta^1_{a,\epsilon}\delta}\right), \tag{17}$$

where Eq. (17) follows from Lemma 8 in [26]. Recall that we have $\mathrm{Reg}^1_\epsilon(T)=\sum_{a:\Delta^1_{a,\epsilon}>0}\Delta^1_{a,\epsilon}N_a(T+1)$. Combining this with the result above, we obtain

$$\mathrm{Reg}_{\epsilon}^{1}(T) \leq \sum_{a:\Delta_{a,\epsilon}^{1}>0} \left(3\Delta_{a,\epsilon}^{1} + \frac{16\Delta_{a,\epsilon}^{1}}{(\Delta_{a,\epsilon}^{1})^{2}} \log \left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\Delta_{a,\epsilon}^{1}\delta} \right) \right). \tag{18}$$

For the second objective, for event UC^{c}, if arm a is selected in round $t \in \mathcal{T}$, then we have $c_a(t) \geq \Delta_a^2/2$ (by Lemma 4). Similar to Eq. (17), this implies that

$$N_a(t) \leq 3 + \frac{16}{(\Delta_a^2)^2} \log \left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\Delta_a^2\delta} \right). \tag{19}$$

In addition, Lemma 5 implies that if arm $a \in \mathcal{S}_2$ is selected in round $t \notin \mathcal{T}$, then $c_a(t) > \epsilon/3$, which implies that $N_a(t) \leq 3 + \frac{36}{\epsilon^2} \log \frac{6e^{\frac{1}{2}}|\mathcal{A}|}{\epsilon\delta}$.

From the two equations above, we observe that for any arm $a \in \mathcal{S}_2$, we have

$$N_a(t) \leq 3 + \frac{16}{(\min\{\Delta_a^2, 2\epsilon/3\})^2} \log \left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\min\{\Delta_a^2, 2\epsilon/3\}\delta} \right). \tag{20}$$

Thus,

$$\mathrm{Reg}^{2}(T) \leq \sum_{a:\Delta_{a}^{2}>0} \left(3\Delta_{a}^{2} + \frac{16\Delta_{a}^{2}}{(\min\{\Delta_a^2, 2\epsilon/3\})^2} \log \left(\frac{4e^{\frac{1}{2}}|\mathcal{A}|}{\min\{\Delta_a^2, 2\epsilon/3\}\delta} \right) \right). \tag{21}$$

Bounds on the expected regret are obtained by using Eq. (12) and setting $\delta = 1/T$. □

The regret bounds given in Theorem 1 are gap-dependent since they are inversely proportional to the suboptimality gaps. This means that the regret is large in problem instances where the suboptimality gaps are small. In contrast to these bounds, the next theorem gives gap-independent regret bounds for ALEX that hold for any problem instance.

Theorem 2 *When ALEX is run with $\delta \in (0,1)$ and $\epsilon > 0$, the following bounds hold with probability at least $1-\delta$ for all $T > 0$:*

$$Reg_{\epsilon}^{1}(T) \leq 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}_1|T} + |\mathcal{S}_1|\Delta_{\max}^{1}, \tag{22}$$

$$Reg^{2}(T) \leq 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}_2|T} + \left(3 + \frac{36}{\epsilon^2} \log \frac{6e^{\frac{1}{2}}|\mathcal{A}|}{\epsilon\delta} \right) |\mathcal{S}_2|\Delta_{\max}^{2}, \tag{23}$$

where $B_{T,\delta} := \sqrt{1 + 2\log(2|\mathcal{A}|T^{1/2}/\delta)}$. Moreover, when ALEX is run with $\delta = 1/T$, we have the following bounds on the expected regret:

$$E[Reg_{\epsilon}^{1}(T)] \leq 4\sqrt{2}B_{T,1/T}\sqrt{|\mathcal{S}_1|T} + (|\mathcal{S}_1| + 1)\Delta_{\max}^{1}, \tag{24}$$

$$E[Reg^{2}(T)] \leq 4\sqrt{2}B_{T,1/T}\sqrt{|\mathcal{S}_2|T} + \left(3|\mathcal{S}_2| + \frac{36|\mathcal{S}_2|}{\epsilon^2} \log \left(\frac{6e^{\frac{1}{2}}|\mathcal{A}|T}{\epsilon} \right) + 1 \right) \Delta_{\max}^{2}. \tag{25}$$

Proof Let $\mathcal{N}_a := \{1 \le t \le T : a(t) = a\}$ and $\tilde{\mathcal{N}}_a := \{t \in \mathcal{N}_a : N_a(t) \ge 1\}$. By Lemma 3, we have for event UC^c (which happens with probability at least $1-\delta$)

$$\mathrm{Reg}_\epsilon^1(T) \le \sum_{a\in\mathcal{S}_1}\sum_{t\in\tilde{\mathcal{N}}_a}(u_a^1(t) - l_a^1(t)) + |\mathcal{S}_1|\Delta_{\max}^1 \tag{26}$$

$$\le 2\sqrt{2}\sum_{a\in\mathcal{S}_1}\left(B_{T,\delta}\sum_{t\in\tilde{\mathcal{N}}_a}\sqrt{\frac{1}{N_a(t)}}\right) + |\mathcal{S}_1|\Delta_{\max}^1 \tag{27}$$

$$\le 4\sqrt{2}B_{T,\delta}\sum_{a\in\mathcal{S}_1}\sqrt{N_a(T)} + |\mathcal{S}_1|\Delta_{\max}^1 \tag{28}$$

$$\le 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}_1|T} + |\mathcal{S}_1|\Delta_{\max}^1, \tag{29}$$

where Eq. (27) holds since $c_a(t) \le \sqrt{2(1+2\log(2|\mathcal{A}|T^{1/2}/\delta))/N_a(t)}$, Eq. (28) follows from the fact that

$$\sum_{k=0}^{N_a(T)-1}\sqrt{\frac{1}{1+k}} \le \int_{x=0}^{N_a(T)}\frac{1}{\sqrt{x}}dx = 2\sqrt{N_a(T)} \tag{30}$$

and Eq. (29) follows from the Cauchy–Schwarz inequality.

The bound for $\mathrm{Reg}^2(T)$ is obtained by using the result in Lemmas 4 and 5. By Lemma 5, we know that

$$\sum_{t\in\mathcal{T}^c}\Delta_{a(t)}^2 \le 3|\mathcal{S}_2|\Delta_{\max}^2 + \frac{36|\mathcal{S}_2|\Delta_{\max}^2}{\epsilon^2}\log\frac{6e^{\frac{1}{2}}|\mathcal{A}|}{\epsilon\delta}. \tag{31}$$

Let $\mathcal{M}_a := \{t \in \mathcal{T} : a(t) = a\}$. Similar to the regret bound proof for objective 1, we have

$$\sum_{t\in\mathcal{T}}\Delta_{a(t)}^2 \le \sum_{a\in\mathcal{S}_2}\sum_{t\in\mathcal{M}_a}(u_a^2(t) - l_a^2(t)) \tag{32}$$

$$\le 2\sqrt{2}\sum_{a\in\mathcal{S}_2}\left(B_{T,\delta}\sum_{t\in\mathcal{M}_a}\sqrt{\frac{1}{N_a(t)}}\right) \tag{33}$$

$$\le 4\sqrt{2}B_{T,\delta}\sqrt{|\mathcal{S}_2|T}. \tag{34}$$

The bound for $\mathrm{Reg}^2(T)$ is obtained by summing the results of Eqs. (31) and (34). Finally, the bounds on the expected regret simply follows from using Eq. (12) and setting $\delta = 1/T$. □

6. Experiments on adaptive multirate multichannel communication

In a cognitive radio network, the SUs are expected to perform under highly dynamic and unpredictable channel conditions by exploiting spatiotemporal spectrum opportunities while avoiding interference with the PUs. Essentially, each SU is required to select a channel that is not currently occupied by a PU, and transmit on that channel with an appropriate rate to maximize its throughput. To accomplish this task, adaptive learning

algorithms that are designed to exploit spectrum opportunities are essential. In the past, MAB algorithms were used for optimal channel and rate selection in cognitive radio networks [4, 10]. Here, we present for the first time, an MAB algorithm for optimal channel and rate selection in a cognitive radio network under a multidimensional performance metric. Essentially, we aim to maximize the SU throughput while ensuring that the PU interference is almost optimal.

6.1. Simulation setup

We consider multirate multichannel communication where the SU selects a transmission rate $r \in \mathcal{R}$ and a channel $c \in \mathcal{C}$ in each round. Here, each transmission rate–channel pair corresponds to an arm. Before transmitting on the selected channel, the SU performs imperfect spectrum sensing with false-positive rate q_{FP} and false-negative rate q_{FN}. We model the PU activity as a Bernoulli random process that is independent over channels and i.i.d. over rounds. Based on this, the probability that the PU is active on channel c is denoted by $q_{PU,c}$, and the PU activity probability vector is given as $\boldsymbol{q}_{PU} = \{q_{PU,c}\}_{c\in\mathcal{C}}$.

The reward in objective 1 is related to PU interference. Basically, the SU receives reward 0 in objective 1 if the PU is present on the channel that it selects but it fails to detect the PU. Otherwise, the reward is 1 in objective 1. The reward in objective 2 is related to SU throughput. If the transmission on the selected channel with the selected rate r is successful, then the reward in objective 2 is $r/r_{\max}$, where $r_{\max}$ is the maximum rate. If there is no transmission or the transmission is unsuccessful (i.e., outage), then the reward in objective 2 is 0. Obviously, the expected reward in objective 2 for rate–channel pair (r, c) depends on the probability of successful transmission, which is given as $1 - p_{\text{out}}(r, c, 1)$ when the PU is present on channel c and $1 - p_{\text{out}}(r, c, 0)$ when the PU is not present on channel c. Here, p_{out} denotes the outage probability, which depends on the rate, the channel gain, the transmit power, and the receiver noise plus interference power.

For every round t and channel c, the transmit power to receiver noise plus interference power ratio $\text{SINR}_{c,t}$ is assumed to be 1 when the PU is not active and is sampled from $\text{Beta}(\alpha, \beta)$ when the PU is active. Thus, the interference caused by the PU presence results in a lower expected $\text{SINR}_{c,t}$. We use Nakagami-m model [27] for channel fading as it captures various fading channels through parameter m. In this model, the gain of channel c in round t, i.e. $h^2_{c,t}$, is gamma distributed with probability density function

$$p(x) = \frac{(\lambda_c m)^m x^{m-1}}{\Gamma(m)} e^{-\lambda_c m x} \tag{35}$$

with shape parameter m, and rate parameter $\lambda_c m$, where $\Gamma(m) := \int_0^\infty t^{m-1} e^{-t} dt$. When $m = 1$, this corresponds to the Rayleigh fading model where the channel gain is exponentially distributed with rate λ_c. The case $0.5 \leq m < 1$ models fading that is more severe than Rayleigh fading and the case $m > 1$ models fading that is less severe than Rayleigh fading. In simulations, we focus on three cases: $m = 0.5$, $m = 1$ and $m = 2$. Based on this, the outage event for rate–channel pair (r, c) is defined as $\log_2(1 + h^2_{c,t}\text{SINR}_{c,t}) < r$.

Parameters used in the simulations are given in Table 2. The given set of parameters corresponds to 9 arms. The expected arm rewards in objectives 1 and 2 are numerically computed by averaging over 5×10^7 random samples, and are given Table 3. Note that the expected reward in objective 1 does not depend on the channel gain. According to this, the best arms in objective 1 are $(2, 2)$, $(1, 2)$ and $(0.5, 2)$ and the lexicographic optimal arm is $(1, 2)$ for $m \in \{0.5, 1, 2\}$. However, for $m = 2$, arm $(0.5, 2)$ is almost as good as arm $(1, 2)$. Existence of multiple best arms in objective 1 is due to the fact that the reward in objective 1 does not depend

on the rate. In all simulations, the time horizon is set to $T = 10^6$ and the reported results correspond to the averages over 50 runs.

Table 2. Simulation parameters. $\boldsymbol{\lambda} = \{\lambda_c\}_{c\in\mathcal{C}}$ denotes the set of channel gain parameters.

$\mathcal{C}$	$\mathcal{R}$	$\boldsymbol{\lambda}$	q_{FP}	q_{FN}	$\boldsymbol{q}_{PU}$	α, β
$\{1,2,3\}$	$\{2,1,0.5\}$	$\{0.5,1,0.5\}$	0.3	0.3	$\{0.2,0.05,0.5\}$	1, 3

Table 3. Expected arm rewards for the simulation parameters given in Table 2.

$a=(r,c)$	(2,1)	(2,2)	(2,3)	(1,1)	(1,2)	(1,3)	(0.5,1)	(0.5,2)	(0.5,3)
μ_a^1 $(m\in\{0.5,1,2\})$	0.940	0.985	0.850	0.940	0.985	0.850	0.940	0.985	0.850
μ_a^2 $(m=0.5)$	0.125	0.055	0.082	0.139	0.106	0.095	0.095	0.087	0.068
μ_a^2 $(m=1)$	0.126	0.033	0.081	0.174	0.123	0.117	0.119	0.111	0.084
μ_a^2 $(m=2)$	0.112	0.012	0.071	0.210	0.135	0.139	0.137	0.134	0.097

6.2. Algorithms

In addition to ALEX, we also report the results of the following algorithms:

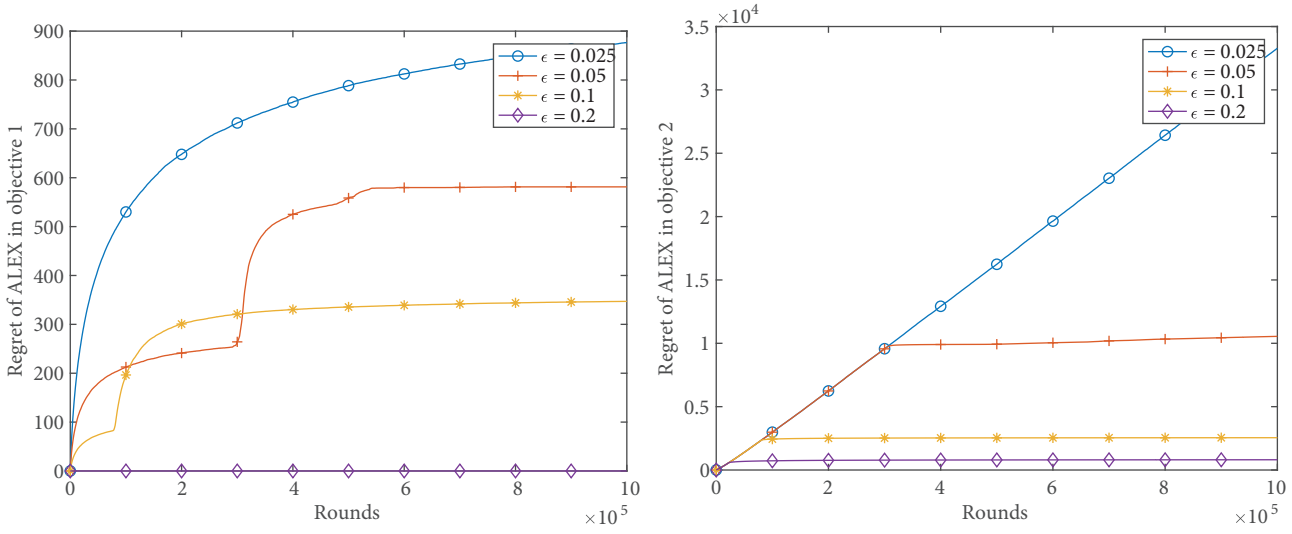
UCB(δ): This is the UCB-based single-objective learning algorithm proposed in [25], which uses a slightly different confidence term than UCB1 in [6] and is proven to achieve bounded regret with high probability. Here, δ denotes the confidence term and is similar to the confidence term of ALEX. In simulations, UCB(δ) learns only from objective 1 and the confidence terms of ALEX and UCB(δ) are set to $\delta = 0.01$.

Empirical Pareto UCB1 (EP-UCB1): This is the UCB-based multiobjective learning algorithm proposed in [15]. This algorithm aims at learning to select arms from the Pareto optimal arm set in order to minimize the Pareto regret. While it is known that a lexicographic optimal arm is also Pareto optimal, the converse does not generally hold [20].

6.3. Results

The regrets of ALEX in objectives 1 and 2 over rounds are shown for different ϵ values for $m = 1$ in Figure 1. Based on this, we conclude that the regret decreases in both objectives as ϵ increases. The regret in objective 1 decreases due to the decreasing suboptimality gaps. Moreover for $\epsilon = 0.1$, 6 out of 9 arms incur no regret in objective 1 and for $\epsilon = 0.2$, all arms incur no regret in objective 1. The regret in objective 2 decreases because for small values of ϵ, ALEX frequently selects the arm with the highest UCB in objective 1 instead of searching for an approximate lexicographic optimal arm in order to make sure that it learns the best arm in objective 1 well. The sharp increase in the regret in objective 1 corresponds to rounds in which ALEX switches its arm selection rule (from line 8 to line 10 in Algorithm 1).

In addition to the regret, the average reward collected by all of the algorithms by the end of the time horizon is given in Table 4. From this, we observe that for ALEX, increasing ϵ decreases the average reward collected in objective 1, while increasing the average reward collected in objective 2 for all values of m. This is expected, since as ϵ increases ALEX makes choices from a larger candidate optimal arm set, which includes arms with higher expected rewards in objective 2 but also lower expected rewards in objective 1. We observe



(a) Objective1 (b) Objective2

Figure 1. Regret of ALEX for different ϵ values in objectives 1 and 2.

Table 4. Average rewards of the algorithms by round T in objectives 1 and 2 respectively.

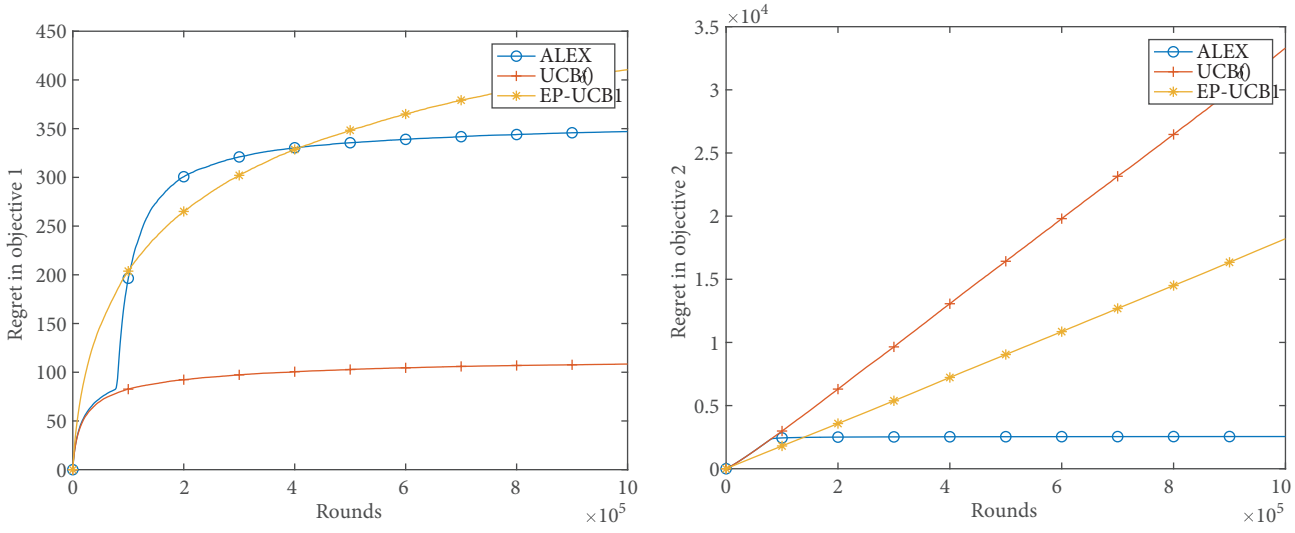
m	ALEX ($\epsilon = 0.025$)	ALEX ($\epsilon = 0.05$)	ALEX ($\epsilon = 0.1$)	ALEX ($\epsilon = 0.2$)	UCB(δ)	EP-UCB1
0.5	0.983, 0.084	0.963, 0.109	0.942, 0.133	0.940, 0.135	0.984, 0.084	0.963, 0.102
1	0.983, 0.090	0.960, 0.123	0.942, 0.167	0.940, 0.171	0.984, 0.090	0.963, 0.115
2	0.983, 0.095	0.965, 0.137	0.942, 0.201	0.940, 0.207	0.984, 0.095	0.965, 0.126

that when $\epsilon = 0.025$, ALEX performs almost the same as UCB(δ), which aims at maximizing the total reward in objective 1. When $\epsilon = 0.2$, the average reward of ALEX in objective 2 is at least 60% higher than that of UCB(δ) and at least 32% higher than that of EP-UCB1, while its average reward in objective 1 is only at most 4.47% lower than that of UCB(δ) and at most 2.59% lower than that of EP-UCB1 for all values of m. These results show the ability of ALEX to tradeoff between the rewards in objectives 1 and 2 by adjusting ϵ.

The regrets of all algorithms are compared in Figure 2 for $m = 1$ and $\epsilon = 0.1$. Note that ϵ does not affect the total reward of UCB(δ) and EP-UCB1 since these algorithms do not take it as input. However, ϵ affects the regrets of these algorithms since it affects the suboptimality gaps of the chosen arms. From the results, we observe that ALEX achieves the smallest regret in objective 2. Moreover, consistent with the theoretical findings, the regret of ALEX exhibits either logarithmic or bounded growth in both objectives, while the regrets of UCB(δ) and EP-UCB1 are linear in objective 2. This shows that UCB(δ) and EP-UCB1 do not have sublinear ϵ lexicographic regret.

Table 5. The fraction of times a 0.1 lexicographic optimal arm is selected.

m	ALEX ($\epsilon = 0.025$)	ALEX ($\epsilon = 0.05$)	ALEX ($\epsilon = 0.1$)	ALEX ($\epsilon = 0.2$)	UCB(δ)	EP-UCB1
0.5	0.343	0.733	0.930	0.956	0.343	0.562
1	0.339	0.660	0.939	0.971	0.341	0.532
2	0.340	0.672	0.945	0.980	0.340	0.589



(a) Objective1 (b) Objective2

Figure 2. Regrets of ALEX, UCB(δ) and EP-UCB1 in objectives 1 and 2 for $\epsilon = 0.1$.

Finally, the fraction of times a 0.1 lexicographic optimal arm is selected is given for all algorithms in Table 5. Results show that ALEX significantly outperforms UCB(δ) and EP-UCB1 in selecting approximate lexicographic optimal arms for $\epsilon = 0.1$ and $\epsilon = 0.2$.

7. Conclusion

In this paper, we proposed a new MAB model called the biobjective MAB and defined the notion of ϵ lexicographic regret. Then, we proposed a learning algorithm called ALEX, and proved that its gap-dependent ϵ lexicographic regret is bounded with high probability and logarithmic in expectation, and its gap-independent regret is $\tilde{O}(\sqrt{T})$ both with high probability and in expectation. Finally, we modeled multirate multichannel communication as a biobjective MAB, and investigated how ALEX learns to tradeoff PU interference and SU throughput better than MAB algorithms that are not tailored to learn approximate lexicographic optimal allocations. Possible future application domains for the biobjective MAB include recommendation engines and robotic systems with multidimensional performance metrics.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant No. 116E229. We thank Robin Ann Downey for proofreading the paper and the anonymous reviewers for their suggestions.

References

[1] Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. Adv Appl Math 1985; 6: 4-22.

[2] Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 1933; 25: 285-294.

[3] Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. In: 19th International Conference on World Wide Web; 26-30 April 2010; Raleigh, NC, USA. New York, NY, USA: ACM. pp. 661-670.

[4] Tekin C, Liu M. Online learning of rested and restless bandits. IEEE Trans Inf Theory 2012; 58: 5588-5611.

[5] Kveton B, Wen Z, Ashkan A, Szepesvari C. Combinatorial cascading bandits. In: 28th Annual Conference on Neural Information Processing Systems; 7-12 December 2015; Montreal, Canada. Red Hook, NY, USA: Curran Associates, Inc. pp. 1450-1458.

[6] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. Mach Learn 2002; 47: 235-256.

[7] Agrawal S, Goyal N. Analysis of Thompson sampling for the multi-armed bandit problem. In: 25th Annual Conference on Learning Theory; 25-27 June 2012; Edinburgh, Scotland. PMLR. pp. 39.1-39.26.

[8] Gai Y, Krishnamachari B, Jain R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. IEEE/ACM Trans Netw 2012; 20: 1466-1478.

[9] Combes R, Proutiere A. Unimodal bandits: regret lower bounds and optimal algorithms. In: 31st International Conference on Machine Learning; 21-16 June 2014; Beijing, China. PMLR. pp. 521-529.

[10] Combes R, Proutiere A. Dynamic rate and channel selection in cognitive radio systems. IEEE J Sel Areas Commun 2015; 33: 910-921.

[11] Gai Y, Krishnamachari B, Jain R. Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation. In: 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum; 6-9 April 2010; Singapore. New York, NY, USA: IEEE. pp. 1-9.

[12] Kveton B, Wen Z, Ashkan A, Szepesvari C. Tight regret bounds for stochastic combinatorial semi-bandits. In: 18th International Conference on Artificial Intelligence and Statistics; 9-12 May 2015; San Diego, CA, USA. PMLR. pp. 535-543.

[13] Chen W, Wang Y, Yuan Y. Combinatorial multi-armed bandit: General framework and applications. In: 30th International Conference on Machine Learning; 16-21 June 2013; Atlanta, GA, USA. PMLR. pp. 151-159.

[14] Chen W, Wang Y, Yuan Y, Wang Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. J Mach Learn Res 2016; 17: 1746-1778.

[15] Drugan MM, Nowé A. Designing multi-objective multi-armed bandits algorithms: a study. In: 2013 International Joint Conference on Neural Networks; 4-9 August 2013; Dallas, TX, USA. New York, NY, USA: IEEE. pp. 1-8.

[16] Drugan MM, Nowé A. Scalarization based Pareto optimal set of arms identification algorithms. In: 2014 International Joint Conference on Neural Networks; 6-11 July 2014; Beijing, China. New York, NY, USA: IEEE. pp. 2690-2697.

[17] Yahyaa SQ, Manderick B. Thompson sampling for multi-objective multi-armed bandits problem. In: 2015 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning; 22-24 April 2015; Brudges, Belgium. Louvain-la-Neuve, Belgium: i6doc.com Publishing. pp. 47-52.

[18] Turgay E, Oner D, Tekin C. Multi-objective contextual bandit problem with similarity information. In: 21st International Conference on Artificial Intelligence and Statistics; 9-11 April 2018; Lanzarote, Spain. PMLR. pp. 1673-1681.

[19] Slivkins A. Contextual bandits with similarity information. J Mach Learn Res 2014; 15: 2533-2568.

[20] Ehrgott M. Multicriteria optimization. 2nd ed. Berlin - Heidelberg, Germany: Springer Science & Business Media, 2005.

[21] Tekin C, Turgay E. Multi-objective contextual bandits with a dominant objective. In: 27th IEEE International Workshop on Machine Learning for Signal Processing; 25-28 September 2017; Tokyo, Japan. New York, NY, USA: IEEE. pp. 1-6.

[22] Tekin C, Turgay E. Multi-objective contextual multi-armed bandit with a dominant objective. IEEE Trans Signal Process 2018; 66: 3799-3813.

[23] Gábor Z, Kalmár Z, Szepesvári C. Multi-criteria reinforcement learning. In: 15th International Conference on Machine Learning; 24-27 July 1998; Madison, WI, USA. San Francisco, CA, USA: Morgan Kaufmann Publishers. pp. 197-205.

[24] Mannor S, Shimkin N. A geometric approach to multi-criterion reinforcement learning. J Mach Learn Res 2004; 5: 325-360.

[25] Abbasi-Yadkori Y, Pál D, Szepesvári C. Improved algorithms for linear stochastic bandits. In: 25th Annual Conference on Neural Information Processing Systems; 12-17 December 2011; Granada, Spain. Red Hook, NY, USA: Curran Associates, Inc. pp. 2312-2320.

[26] Antos A, Grover V, Szepesvári C. Active learning in heteroscedastic noise. Theor Comput Sci 2010; 411: 2712-2728.

[27] Stuber GL. Principles of mobile communication. 2nd ed. Norwell, MA, USA: Kluwer, 2001.