2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

MULTI-OBJECTIVE CONTEXTUAL BANDITS WITH A DOMINANT OBJECTIVE

Cem Tekin, Eralp Turgay

Bilkent University
Department of Electrical and Electronics Engineering
Ankara, Turkey

ABSTRACT

In this paper, we propose a new contextual bandit problem with two objectives, where one of the objectives dominates the other objective. Unlike single-objective bandit problems in which the learner obtains a random scalar reward for each arm it selects, in the proposed problem, the learner obtains a random reward vector, where each component of the reward vector corresponds to one of the objectives. The goal of the learner is to maximize its total reward in the non-dominant objective while ensuring that it maximizes its reward in the dominant objective. In this case, the optimal arm given a context is the one that maximizes the expected reward in the non-dominant objective among all arms that maximize the expected reward in the dominant objective. For this problem, we propose the multi-objective contextual multi-armed bandit algorithm (MOC-MAB), and prove that it achieves sublinear regret with respect to the optimal context dependent policy. Then, we compare the performance of the proposed algorithm with other state-of-the-art bandit algorithms. The proposed contextual bandit model and the algorithm have a wide range of real-world applications that involve multiple and possibly conflicting objectives ranging from wireless communication to medical diagnosis and recommender systems.

Index Terms— Online learning, contextual bandits, multi-objective bandits, dominant objective, regret bounds.

1. INTRODUCTION

With the rapid increase in the generation speed of the streaming data, online learning methods are becoming increasingly valuable for sequential decision making problems. Many of these problems, ranging from recommender systems [1] to medical screening and diagnosis [2, 3] to cognitive radio networks [4] involve multiple and possibly conflicting objectives. In this work, we propose a multi-objective contextual bandit problem with dominant and non-dominant objectives. For this problem, we construct a multi-objective contextual bandit algorithm named MOC-MAB, which maximizes long-term reward of the non-dominant objective conditioned on the fact that it maximizes the long-term reward of the dominant objective.

In this problem, the learner observes a multi-dimensional context vector in each time step. Then, it selects one of the available arms and receives a random reward for each objective, which is drawn from a fixed distribution that depends on the context and the selected arm. No statistical assumptions are made on the way the contexts arrive, and the learner does not have any a priori information on the reward distributions. The optimal arm is defined as the one that maximizes the expected reward of the non-dominant objective among all arms that maximizes the expected reward of the dominant objective given the context vector.

The learner's performance is measured in terms of its regret, which is the difference between the expected total reward of an oracle that knows the optimal arm given each context and that of the learner. We prove that MOC-MAB achieves $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ regret in both objectives, where d is the dimension of the context vector and α is a constant that depends on the similarity information that relates the distances between contexts to the distances between expected rewards of an arm. This shows that MOC-MAB is average-reward optimal in the limit $T \to \infty$. In addition, we also evaluate the performance of MOC-MAB through simulations and compare it with other single-objective and multi-objective bandit algorithms. Our results show that MOC-MAB outperforms its competitors, which are not specifically designed to deal with problems involving dominant and non-dominant objectives.

2. RELATED WORK

Multi-objective bandits [5] and contextual bandits [6, 7, 8] are two different extensions of the classical multi-armed bandit problem [9], which have been studied extensively but separately.

Existing works on contextual bandits can be categorized into three. The first category assumes the existence of similarity information (usually provided in terms of a metric) that relates the variation in the expected reward of an arm as a function of the context to the distance between the contexts. This problem is considered in [10], and a learning algorithm that

978-1-5090-6341-3/17/$31.00

achieves sublinear in time regret is proposed. The main idea is to partition the context space and to estimate the expected arm rewards for each set in the partition separately. In [11], it is assumed that the arm rewards depend on an unknown subset of the contexts, and it is shown that the regret in this case only depends on the number of relevant context dimensions. For this category, no statistical assumptions are made on the context arrivals. The second category assumes that the expected reward of an arm is a linear combination of the elements of the context vector. For this model, Li et al. [1] proposed the LinUCB algorithm. A modified version of this algorithm, named SupLinUCB [12], is shown to achieve $\tilde{O}(\sqrt{Td})$ regret, where d is the dimension of the context vector. The third category assumes that the contexts and arm rewards are drawn from a fixed but unknown distribution. For this case, Langford et al. proposed the epoch greedy algorithm with $O(T^{2/3})$ regret and later works [13, 14] proposed more efficient learning algorithms with $\tilde{O}(T^{1/2})$ regret. Our problem is similar to the problems in the first category in terms of the context arrivals and existence of the similarity information.

Similarly, existing works on multi-objective bandits can be categorized into two: Pareto approach and scalarized approach. In the Pareto approach, the main idea is to estimate the Pareto front set which consists of the arms that are not dominated by any other arm. Dominance relationship is defined such that if expected reward of an arm a^* is greater than expected reward of another arm a in at least one objective, and expected reward of the arm a is not greater than expected reward of the arm a^* in any objective, then the arm a^* dominates the arm a. For instance, in [5] learning algorithms that compute upper confidence bounds (UCBs) as the index for each objective, use these indices to compute the Pareto front set, and select arms randomly from the Pareto front set are proposed. Numerous other algorithms are also proposed in prior works, including the Pareto Thompson sampling algorithm in [15] and the Annealing Pareto algorithm in [16]. On the other hand, in the scalarized approach [5, 17], a random weight is assigned to each objective at each time step, from which a weighted sum of indices of the objectives are calculated. The regret notion used in Pareto and scalarized approaches are very different from our regret notion. In the Pareto approach, the regret at time step t is defined as the minimum distance that should be added to expected reward vector of the chosen arm at time t to move the chosen arm to the Pareto front set. On the other hand, scalarized regret is the difference between scalarized expected rewards of the optimal arm and the chosen arm.

3. PROBLEM DESCRIPTION

The system operates in a sequence of discrete time steps indexed by $t \in \{1, 2, \ldots\}$. At the beginning of time step t, the learner observes a d-dimensional context vector denoted by x_t. Without loss of generality, we assume that x_t lies in the context space $\mathcal{X} := [0, 1]^d$. After observing x_t, the learner selects an arm a_t from a finite set $\mathcal{A}$. Then, the learner observes a two dimensional random reward $\boldsymbol{r}_t = (r_t^1, r_t^2)$, which is drawn from a fixed probability distribution that depends on both x_t and a_t, and has support in $[0, 1]^2$. This distribution is unknown to the learner. Here, r_t^1, r_t^2 denotes the rewards in the dominant and non-dominant objectives, respectively.

The expected rewards for the dominant and non-dominant objectives for context-arm pair (x, a) are denoted by $\mu_a^1(x)$ and $\mu_a^2(x)$, respectively. Hence, the random rewards can be written as $r_t^1 = \mu_{a_t}^1(x_t) + \kappa_t^1$, $r_t^2 = \mu_{a_t}^2(x_t) + \kappa_t^2$, where the noise process $\{(\kappa_t^1, \kappa_t^2)\}$ is such that the marginal distribution of κ_t^i is conditionally 1-sub-Gaussian, i.e., $\forall \lambda \in \mathbb{R}$, $\mathrm{E}[e^{\lambda \kappa_t^i} | \boldsymbol{a}_{1:t}, \boldsymbol{\kappa}_{1:t-1}^1, \boldsymbol{\kappa}_{1:t-1}^2, x_{1:t}] \leq \exp(\lambda^2/2)$ where $\boldsymbol{b}_{1:t} := (b_1, \ldots, b_t)$. The set of arms that maximize the expected reward for the dominant objective for context x is given as $\mathcal{A}^*(x) := \arg\max_{a \in \mathcal{A}} \mu_a^1(x)$. The set of optimal arms is given as the set of arms in $\mathcal{A}^*(x)$ with the highest expected rewards for the non-dominant objective. Without loss of generality, we assume that there is a single optimal arm, and denote it by $a^*(x)$. Hence, we have $a^*(x) = \arg\max_{a \in \mathcal{A}^*(x)} \mu_a^2(x)$. Let $\mu_*^1(x)$ and $\mu_*^2(x)$ denote the expected rewards of arm $a^*(x)$ in the dominant and the non-dominant objectives, respectively, when the context is x. We assume that the expected rewards are Hölder continuous in the context.

Assumption 1. *There exists $L > 0, \alpha > 0$ such that for all $i \in \{1, 2\}$, $a \in \mathcal{A}$ and $x \in \mathcal{X}$, we have $|\mu_a^i(x) - \mu_a^i(x')| < L \|x - x'\|^\alpha$.*

Initially, the learner does not know the expected rewards; it learns them over time. The goal of the learner is to compete with an oracle, which knows the expected rewards of the arms for every context and chooses the optimal arm given the current context. The multi-objective regret of the learner by time step T is defined as the tuple $(\mathrm{Reg}^1(T), \mathrm{Reg}^2(T))$, where

$$\mathrm{Reg}^i(T) := \sum_{t=1}^{T} \mu_*^i(x_t) - \sum_{t=1}^{T} \mu_{a_t}^i(x_t), \; i \in \{1, 2\} \qquad (1)$$

for an arbitrary sequence of contexts $x_1, \ldots, x_T$. Two real-world applications of the proposed contextual bandit model are given below.

Multi-channel Communication: Consider a multi-channel communication scenario in which a user chooses a channel $Q \in \mathcal{Q}$ and a transmission rate $R \in \mathcal{R}$ in each time step after receiving context $x_t := \{x_{Q,t}\}_{Q \in \mathcal{Q}}$, where $x_{Q,t}$ is the noise and interference level on channel Q in time step t. In this setup, each arm corresponds to a transmission rate-channel pair denoted by $a_{R,Q}$. Hence, the set of arms is $\mathcal{A} = \mathcal{R} \times \mathcal{Q}$. When the user completes its transmission at the end of time step t, it receives a two dimensional reward where the dominant one is related to throughput and the non-dominant one is related to reliability. Here, $r_t^2 \in \{0, 1\}$

where 0 and 1 correspond to failed and successful transmission, respectively. Moreover, the success probability of $a_{R,Q}$ is equal to $\mu^2_{a_{R,Q}}(x_t) = 1 - p_{\text{out}}(R, Q, x_t)$, where $p_{\text{out}}(\cdot)$ denotes the outage probability. Here, $p_{\text{out}}(R, Q, x_t)$ also depends on the gain on channel Q whose distribution is unknown to the user. On the other hand, for $a_{R,Q}$, $r^1_t \in \{0, R\}$ and $\mu^1_{a_{R,Q}}(x_t) = R(1 - p_{\text{out}}(R, Q, x_t))$. It is usually the case that the outage probability increases with R, so maximizing the throughput and reliability are usually conflicting objectives.

Online Binary Classification: Consider a medical diagnosis problem where a patient with context x_t (including features such as age, gender, medical test results etc.) arrives in time step t. Then, this patient is assigned to one of the experts in $\mathcal{A}$ who will diagnose the patient. In reality, these experts can either be clinical decision support systems or humans, but the classification performance of these experts are context dependent and unknown a priori. In this problem, the dominant objective can correspond to accuracy while the non-dominant objective can correspond to false negative rate. For this case, the rewards in both objectives are binary, and depend on whether the classification is correct or a positive case is correctly identified.

4. THE LEARNING ALGORITHM AND ITS REGRET

The pseudocode of MOC-MAB is given in Algorithm 1. MOC-MAB uniformly partitions $\mathcal{X}$ into m^d hypercubes with edge lengths $1/m$. This partition is denoted by $\mathcal{P}$. For each $p \in \mathcal{P}$ and $a \in \mathcal{A}$ it keeps: (i) a counter $N_{a,p}$ that counts the number of times arm a is selected when the context arrived to p, (ii) the sample mean of the rewards obtained from selections of arm a when the contexts is in p, i.e., $\hat{\mu}^1_{a,p}$ and $\hat{\mu}^2_{a,p}$ for the dominant and non-dominant objectives, respectively.

At time step t, MOC-MAB first identifies the hypercube in $\mathcal{P}$ that contains x_t, which is denoted by p^*. Then, it calculates the following UCBs for the rewards in dominant and non-dominant objectives:

$$g^i_{a,p^*} := \hat{\mu}^i_{a,p^*} + u_{a,p^*},\ i \in \{1, 2\} \tag{2}$$

where the uncertainty level $u_{a,p} := \sqrt{2A_{m,T}/N_{a,p}}$, $A_{m,T} := (1 + 2\log(4|\mathcal{A}|m^d T^{3/2}))$ represents the uncertainty over the sample mean estimate of the reward due to the number of instances that are used to compute $\hat{\mu}^i_{a,p^*}$. Hence, a UCB for $\mu^i_a(x)$ is $g^i_{a,p} + v$ for $x \in p$, where $v := Ld^{\alpha/2}m^{-\alpha}$ denotes the *margin of tolerance* due to the partitioning of $\mathcal{X}$. The main learning principle in such a setting is called optimism under the face of uncertainty. The idea is to inflate the reward estimates from arms that are not selected often by a certain level, such that the inflated reward estimate becomes an upper confidence bound for the true expected reward with a very high probability. This way, arms that are not selected frequently are explored, and this exploration potentially helps the learner to discover arms that are better than the arm with the highest estimated reward. As expected, the uncertainty level vanishes as an arm gets selected more often. Then, MOC-MAB judiciously determines the arm to select based on these UCBs. It is important to note that the choice $a^*_1 := \arg\max_{a\in\mathcal{A}} g^1_{a,p^*}$ can be highly suboptimal for the non-dominant objective. To see this, consider a very simple setting, where $\mathcal{A} = \{a, b\}$, $\mu^1_a(x) = \mu^1_b(x) = 0.5$, $\mu^2_a(x) = 1$ and $\mu^2_b(x) = 0$ for all $x \in \mathcal{X}$. For an algorithm for which $a_t = a^*_1$ always, both arms will be equally selected in expectation (assuming that the ties are randomly broken). Hence, due to the noisy rewards, arm 2 will be selected more than half of the time with some non-zero probability. For each such sample path, the regret in the non-dominant objective is linear in T. This implies that the expected regret is also linear in T. MOC-MAB overcomes the effect of the noise mentioned above due to the randomness in the rewards and the partitioning of $\mathcal{X}$ by creating a safety margin below the maximal index $g^1_{a^*_1,p^*}$ for the dominant objective, when its confidence for a^*_1 is high, i.e., when $u_{a^*_1,p^*} \leq \beta v$, where $\beta > 0$ is a constant. For this, it calculates the set of candidate optimal arms given as

$$\hat{\mathcal{A}}^* := \left\{a \in \mathcal{A} : g^1_{a,p*} \geq \hat{\mu}^1_{a^*_1,p^*} - u_{a^*_1,p^*} - 2v\right\}. \tag{3}$$

Then, it selects $a_t \in \arg\max_{a\in\hat{\mathcal{A}}^*} g^2_{a,p*}$. On the other hand, when its confidence for a^*_1 is low, i.e., when $u_{a^*_1,p^*} > \beta v$, it has a little hope even in selecting an optimal arm for the dominant objective. In this case it just selects $a_t = a^*_1$ to improve its confidence for a^*_1. After its arm selection, it receives the random reward vector $\boldsymbol{r}_t$, which is then used to update the counters and the sample mean rewards for p^*. The above procedure repeats at every time step t.

The following theorem bounds the expected regret of MOC-MAB.

Theorem 1. *When MOC-MAB is run with inputs* $m = \lceil T^{1/(3\alpha+d)} \rceil$ *and* $\beta > 0$*, we have*

$$\begin{aligned}\mathrm{E}[Reg^1(T)] \leq & C^1_{\max} + 2^d|\mathcal{A}|C^1_{\max}T^{\frac{d}{3\alpha+d}} \\ & +2(\beta+2)Ld^{\alpha/2}T^{\frac{2\alpha+d}{3\alpha+d}} \\ & +2^{d/2+1}B_{m,T}\sqrt{|\mathcal{A}|}T^{\frac{1.5\alpha+d}{3\alpha+d}}\end{aligned}$$

$$\begin{aligned}\mathrm{E}[Reg^2(T)] \leq & \ 2^{d/2+1}B_{m,T}\sqrt{|\mathcal{A}|}T^{\frac{1.5\alpha+d}{3\alpha+d}} + C^2_{\max} \\ & + \left(2Ld^{\alpha/2} + \frac{C^2_{\max}|\mathcal{A}|2^{1+2\alpha+d}A_{m,T}}{\beta^2L^2d^\alpha}\right)T^{\frac{2\alpha+d}{3\alpha+d}} \\ & +2^dC^2_{\max}|\mathcal{A}|T^{\frac{d}{3\alpha+d}}\end{aligned}$$

where $C^i_{\max}$ *is the maximum difference between expected rewards of the arms in objective* i*, and* $B_{m,T} := 2\sqrt{2A_{m,T}}$.

It can be shown that when we set $m = \lceil T^{1/(2\alpha+d)} \rceil$ regret bound of the dominant objective becomes $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$. However, in this case the regret bound of the non-dominant objective becomes $O(T)$. Hence, the value for m that makes the time order of both regrets equal is $m = \lceil T^{1/(3\alpha+d)} \rceil$.

Algorithm 1 MOC-MAB

Input: T, d, L, α, m, β
Initialize sets: Create partition $\mathcal{P}$ of $\mathcal{X}$ into m^d identical hypercubes
Initialize counters: $N_{a,p} = 0, \forall a \in \mathcal{A}, \forall p \in \mathcal{P}, t = 1$
Initialize estimates: $\hat{\mu}^1_{a,p} = \hat{\mu}^2_{a,p} = 0, \forall a \in \mathcal{A}, \forall p \in \mathcal{P}$
while $1 \leq t \leq T$ **do**
  Find $p^* \in \mathcal{P}$ such that $x_t \in p^*$
  Compute g^i_{a,p^*} for $a \in \mathcal{A}$, $i \in \{1, 2\}$ as given in (2)
  Set $a^*_1 \in \arg\max_{a \in A} g^1_{a,p^*}$.
  if $u_{a^*_1,p^*} > \beta v$ **then**
    Select arm $a_t = a^*_1$
  else
    Find set of candidate optimal arms $\hat{\mathcal{A}}^*$ given in (3)
    Select arm $a_t \in \arg\max_{a \in \hat{\mathcal{A}}^*} g^2_{a,p^*}$
  end if
  Observe $\boldsymbol{r}_t = (r^1_t, r^2_t)$
  $\hat{\mu}^i_{a_t,p^*} \leftarrow (\hat{\mu}^i_{a_t,p^*} N_{a_t,p^*} + r^i_t)/(N_{a_t,p^*} + 1), i \in \{1, 2\}$
  $N_{a_t,p^*} \leftarrow N_{a_t,p^*} + 1$
  $t \leftarrow t + 1$
end while

5. ILLUSTRATIVE RESULTS

In this section, we numerically evaluate the performance MOC-MAB on a synthetic multi-objective dataset, and compare it with other bandit algorithms. We take $\mathcal{X} = [0, 1]^2$ and assume that the context at each time step is chosen uniformly at random from $\mathcal{X}$. We assume that there are 3 arms and $T = 100000$. The expected arm rewards are generated as follows. We generate three multivariate Gaussian distributions both for the dominant and non-dominant objectives. For the dominant objective, the mean vectors of the first two distributions are $[0.35, 0.5]$ and the mean vector of the third distribution is $[0.65, 0.5]$. Similarly, for the non-dominant objective, the mean vectors of the distributions are $[0.35, 0.65]$, $[0.35, 0.35]$ and $[0.65, 0.5]$, respectively. For all the distributions the covariance matrix is given by $0.3 * \mathrm{I}$ where I is the 2 by 2 identity matrix. Then, each Gaussian distribution is normalized by multiplying the distribution with a constant, such that its maximum value becomes 1. These normalized Gaussian distributions form the expected arm rewards. We assume that the random reward of an arm in an objective given a context x is a Bernoulli random variable whose parameter is equal to the magnitude of the corresponding normalized Gaussian distribution at context x.

We compare MOC-MAB with the following algorithms:
Pareto UCB1 (P-UCB1): This is the Empirical Pareto UCB1 algorithm proposed in [5].
Scalarized UCB1 (S-UCB1): This is the Scalarized Multi-objective UCB1 algorithm proposed in [5].
Contextual Pareto UCB1 (CP-UCB1): This is the contextual version of P-UCB1 which partitions the context space in the same way as MOC-MAB does, and uses a different instance of P-UCB1 in each set of the partition.
Contextual Scalarized UCB1 (CS-UCB1): This is the contextual version of S-UCB1, which partitions the context space in the same way as MOC-MAB does, and uses a different instance of S-UCB1 in each set of the partition.
Contextual Dominant UCB1 (CD-UCB1): This is the contextual version of UCB1 [18], which partitions the context space in the same way as MOC-MAB does, and uses a different instance of UCB1 in each set of the partition. This algorithm only uses the rewards from the dominant objective to update the indices of the arms.

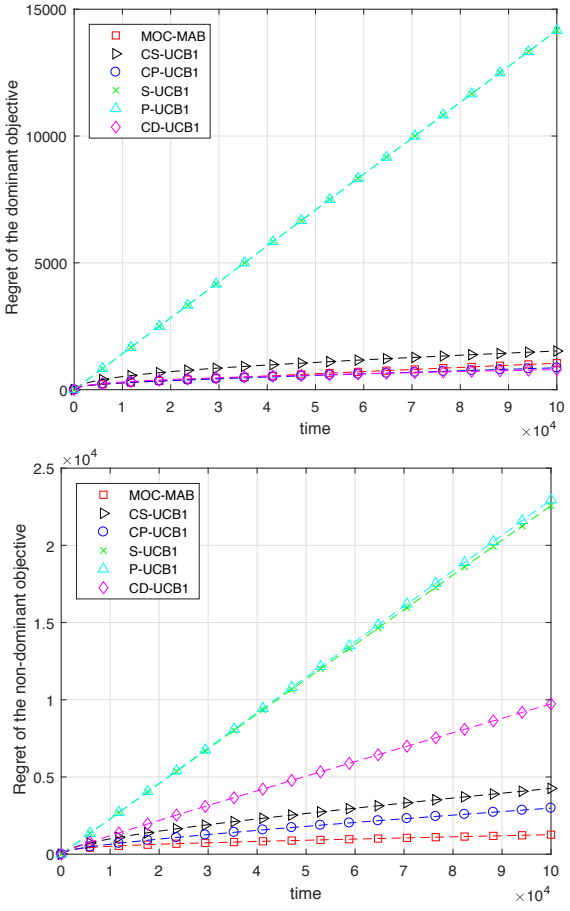

Fig. 1. Regrets of MOC-MAB and the other algorithms.

For S-UCB1 and CS-UCB1, the weights of the linear scalarization functions are chosen as $[1, 0]$, $[0.5, 0.5]$ and $[0, 1]$. For all contextual algorithms, the partition of the context space is formed by choosing m according to Theorem 1. For MOC-MAB, β is chosen as 0.1. In addition, we scaled down the uncertainty level (also known as the inflation term) of all the algorithms by a constant chosen from $\{1, 1/5, 1/10, 1/15, 1/20, 1/25\}$, since we observed that the regrets of the algorithms become smaller when the uncertainty level is scaled down. For MOC-MAB the optimal scale

factor for the dominant objective is $1/20$, for CP-UCB1 and S-UCB1, it is $1/10$, for CS-UCB1 and CD-UCB1, it is $1/5$ and for P-UCB1, it is 1. The regret results are obtained by using the optimal scale factor for each algorithm. Every algorithm is run 1000 times and the results are averaged over these runs. Simulation results given in Fig. 1 show the change in the regret of the algorithms in both objectives as a function of time. As observed from the results, MOC-MAB beats all other algorihms in both objectives except CD-UCB1 and CP-UCB1. While the regrets of these algorithms in the dominant objective are slightly better than that of MOC-MAB, their regrets are much worse than MOC-MAB in the non-dominant objective. The total reward of MOC-MAB in the dominant objective is 0.8% higher than that of CS-UCB1, and 29.7% higher than that of non-contextual algoritms but 0.4% smaller than that of CD-UCB1 and 0.3% smaller than that of CP-UCB1. In the non-dominant objective, total reward of MOC-MAB is 2.7% higher than that of CP-UCB1, 4.8% higher than that of CS-UCB1, 15% higher than that of CD-UCB1, 49.1% higher than that of S-UCB1, and 50.5% higher than that of P-UCB1.

6. PROOF OF THEOREM 1

For all the parameters defined in Section 4, we explicitly use the time index t, when referring to the value of that parameter at the beginning of time step t. For instance, $N_{a,p}(t)$ denotes the value of $N_{a,p}$ at the beginning of time step t. Let $N_p(t)$, denote the number of context arrivals to $p \in \mathcal{P}$ by time step t, $\tau_p(t)$ denote the time step in which a context arrives to $p \in \mathcal{P}$ for the tth time, and $R_a^i(t)$ denote the random reward of arm a in objective i at time step t. Let $\tilde{x}_p(t) := x_{\tau_p(t)}$, $\tilde{R}_{a,p}^i(t) := R_a^i(\tau_p(t))$, $\tilde{N}_{a,p}(t) := N_{a,p}(\tau_p(t))$, $\tilde{\mu}_{a,p}^i(t) := \hat{\mu}_{a,p}^i(\tau_p(t))$, $\tilde{a}_p(t) := a_{\tau_p(t)}$, and $\tilde{u}_{a,p}(t) := u_{a,p}(\tau_p(t))$. Next, we define the following lower and upper bounds: $L_{a,p}^i(t) := \tilde{\mu}_{a,p}^i(t) - \tilde{u}_{a,p}(t)$ and $U_{a,p}^i(t) := \tilde{\mu}_{a,p}^i(t) + \tilde{u}_{a,p}(t)$ for $i \in \{1,2\}$. Let $\text{UC}_{a,p}^i := \bigcup_{t=1}^{N_p(T)} \{\mu_a^i(\tilde{x}_p(t)) \notin [L_{a,p}^i(t) - v, U_{a,p}^i(t) + v]\}$ denote the event that the learner is not confident about its reward estimate in objective i for at least once in time steps in which the contexts is in p by time T. Also, let $\text{UC}_p^i := \cup_{a \in \mathcal{A}} \text{UC}_{a,p}^i$, $\text{UC}_p := \cup_{i \in \{1,2\}} \text{UC}_p^i$ and $\text{UC} := \cup_{p \in \mathcal{P}} \text{UC}_p$.

Lemma 1. $\Pr(UC) \leq 1/T$.

Proof. (Sketch) From the definitions of $L_{a,p}^i(t)$, $U_{a,p}^i(t)$ and $\text{UC}_{a,p}^i$, it is clear that $\text{UC}_{a,p}^i$ does not happen when $\tilde{\mu}_{a,p}^i(t)$ remains close to $\mu_a^i(\tilde{x}_p(t))$ for all $t \in \{1, \ldots, N_p(T)\}$. This motivates us to use the concentration inequality given in Lemma 6 in [19] to bound the probability of $\text{UC}_{a,p}^i$. However, a direct application of this inequality is not possible to our problem, due to the fact that the context sequence $\tilde{x}_p(1), \ldots, \tilde{x}_p(N_p(t))$ does not have identical elements, which makes the expected values of $\tilde{R}_{a,p}^i(1), \ldots, \tilde{R}_{a,p}^i(N_p(t))$ different. To overcome this problem, we define two new sequences of random variables that upper and lower bound the sequence of random variables $\tilde{R}_{a,p}^i(1), \ldots, \tilde{R}_{a,p}^i(N_p(t))$ for each t. We call the sequence that lower bounds our sequence as the *worst sequence* and the sequence that upper bounds our sequence as *best sequence*. Then, we show that $\text{UC}_{a,p}^i$ is included in the event that the sample mean estimate of the rewards from either the worst sequence or the best sequence do not lie between the lower and upper confidence bounds for at least one t. Finally, we apply Lemma 6 in [19] to the worst and best sequence, and then apply a union bound to bound $\Pr(\text{UC})$. □

Using Lemma 1 and the law of total expectation, we obtain

$$\text{E}[\text{Reg}^i(T)] \leq C_{\max}^i + \text{E}[\text{Reg}^i(T)|\text{UC}^c]. \tag{4}$$

We bound $\text{E}[\text{Reg}^i(T)|\text{UC}^c]$ in the rest of the proof. For the simplicity of notation we let $a^*(t) := a^*(\tilde{x}_p(t))$ denote the optimal arm, $\tilde{a}(t) := \tilde{a}_p(t)$ denote the selected arm and $\hat{a}_1^*(t)$ denote the arm whose index for the dominant objective is the highest at time $\tau_p(t)$. It can be shown that on event UC^c, we have

$$\begin{aligned}&\mu_{a^*(t)}^1(\tilde{x}_p(t)) - \mu_{\tilde{a}(t)}^1(\tilde{x}_p(t)) \leq U_{\tilde{a}(t),p}^1(t) \\ &- L_{\tilde{a}(t),p}^1(t) + 2(\beta+2)v, \ \ \forall t \in \{1, \ldots, N_p(T)\}.\end{aligned} \tag{5}$$

Moreover, when $\tilde{u}_{\hat{a}_1^*(t),p}(t) \leq \beta v$ holds under UC^c, we also have

$$\begin{aligned}&\mu_{a^*(t)}^2(\tilde{x}_p(t)) - \mu_{\tilde{a}(t)}^2(\tilde{x}_p(t)) \\ &\leq U_{\tilde{a}(t),p}^2(t) - L_{\tilde{a}(t),p}^2(t) + 2v.\end{aligned} \tag{6}$$

(5) and (6) allows us to bound the regrets at time step $\tau_p(t)$ in terms of the gap between the upper and lower confidence bounds, which we expect to shrink as an arm gets selected. However, the term with v, which appears due to the partitioning of the context space does not change as the number of observations increase.

We obtain the bound for $\text{E}[\text{Reg}^1(T)|\text{UC}^c]$, by simply summing (5) over all time steps and taking the expectation. For this, we let $\text{Reg}_p^i(T) := \sum_{t=1}^{N_p(T)} \mu_*^i(\tilde{x}_p(t)) - \sum_{t=1}^{N_p(T)} \mu_{\tilde{a}_p(t)}^i(\tilde{x}_p(t))$ denote the regret in objective i that is incurred in time steps when the context is in $p \in \mathcal{P}$. For $\mathcal{T}_{a,p} := \{t \leq N_p(T) : \tilde{a}_p(t) = a\}$ and $\tilde{\mathcal{T}}_{a,p} := \{t \in \mathcal{T}_{a,p} : \tilde{N}_{a,p}(t) \geq 1\}$ it can be shown that

$$\begin{aligned}&\text{E}[\text{Reg}_p^1(T)|\text{UC}^c] \leq |\mathcal{A}|C_{\max}^1 + 2(\beta+2)vN_p(T) \\ &+ \text{E}[\sum_{a \in \mathcal{A}} \sum_{t \in \tilde{\mathcal{T}}_{a,p}} U_{\tilde{a}_p(t),p}^1(t) - L_{\tilde{a}_p(t),p}^1(t)|\text{UC}^c] \\ &\leq |\mathcal{A}|C_{\max}^1 + 2(\beta+2)vN_p(T) + 4\sqrt{2A_{m,T}|\mathcal{A}|N_p(T)}.\end{aligned}$$

By summing the above term over all $p \in \mathcal{P}$, we obtain the bound for $\mathrm{E}[\mathrm{Reg}^1(T)|\mathrm{UC}^c]$. In order to obtain the bound for $\mathrm{E}[\mathrm{Reg}^2(T)|\mathrm{UC}^c]$, we also need to take into account the time steps for which $\tilde{u}_{\hat{a}_1^*(t),p}(t) > \beta v$. It can be shown that the number of such time steps is bounded by $|\mathcal{A}|(2A_{m,T}/(\beta v)^2 + 1)$. The expected regret in the non-dominant objective for each of these steps is bounded by $C_{\max}^2$. Then, using the same technique as in bounding $\mathrm{E}[\mathrm{Reg}_p^1(T)|\mathrm{UC}^c]$, we obtain

$$\mathrm{E}[\mathrm{Reg}_p^2(T)|\mathrm{UC}^c] \leq C_{\max}^2|\mathcal{A}|(2A_{m,T}/(\beta v)^2 + 1) + 2vN_p(T) + 4\sqrt{2A_{m,T}|\mathcal{A}|N_p(T)}.$$

By summing the above term over all $p \in \mathcal{P}$, we obtain the bound for $\mathrm{E}[\mathrm{Reg}^2(T)|\mathrm{UC}^c]$. Then, we substitute the results in (4) to bound the expected regrets in both objectives. The resulting bound depend on parameter m. Our aim is to chose m such that the time order of the growth rate of the regret in both objectives is balanced, which is achieved by taking $m = \lceil T^{1/(3\alpha+d)} \rceil$.

7. ACKNOWLEDGEMENT

This work is supported by TUBITAK 2232 Grant 116C043 and supported in part by TUBITAK 3501 Grant 116E229.

8. REFERENCES

[1] Lihong Li, Wei Chu, John Langford, and Robert E Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 661–670.

[2] Linqi Song, William Hsu, Jie Xu, and Mihaela van der Schaar, "Using contextual learning to improve diagnostic accuracy: Application in breast cancer screening," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 3, pp. 902–914, 2016.

[3] Cem Tekin, Jinsung Yoon, and Mihaela van der Schaar, "Adaptive ensemble learning with confidence bounds," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 888–903, 2017.

[4] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. Symp. Dynamic Spectrum Access Networks (DySPAN)*, 2010, pp. 1–9.

[5] Madalina M Drugan and Ann Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *Proc. Int. Joint Conf. Neural Networks*, 2013, pp. 1–8.

[6] John Langford and Tong Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. NIPS*, 2007, vol. 20, pp. 1096–1103.

[7] Aleksandrs Slivkins, "Contextual bandits with similarity information," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2533–2568, 2014.

[8] Cem Tekin and Mihaela van der Schaar, "Distributed online learning via cooperative contextual bandits," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3700–3714, 2015.

[9] Tze L Lai and Herbert Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[10] Tyler Lu, Dávid Pál, and Martin Pál, "Contextual multi-armed bandits," in *Proc. AISTATS*, 2010, pp. 485–492.

[11] Cem Tekin and Mihaela van der Schaar, "RELEAF: An algorithm for learning and exploiting relevance," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 716–727, 2015.

[12] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire, "Contextual bandits with linear payoff functions," in *Proc. AISTATS*, 2011.

[13] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang, "Efficient optimal learning for contextual bandits," in *Proc. UAI*, pp. 169–178.

[14] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *Proc. ICML*, 2014, pp. 1638–1646.

[15] Saba Q Yahyaa and Bernard Manderick, "Thompson sampling for multi-objective multi-armed bandits problem," in *Proc. 23rd European Symp. Artificial Neural Networks (ESANN)*, 2015, pp. 47–52.

[16] Saba Q Yahyaa, Madalina M Drugan, and Bernard Manderick, "Annealing-Pareto multi-objective multi-armed bandit algorithm," in *Proc. Symp. Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2014, pp. 1–8.

[17] Madalina M Drugan and Ann Nowé, "Scalarization based Pareto optimal set of arms identification algorithms," in *Proc. Int. Joint Conf. Neural Networks*, 2014, pp. 2690–2697.

[18] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[19] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. NIPS*, 2011, vol. 24, pp. 2312–2320.