Multi-objective Contextual Multi-armed Bandit With a Dominant Objective

Cem Tekin, *Member, IEEE*, and Eralp Turğay

***Abstract*—We propose a new multi-objective contextual multi-armed bandit (MAB) problem with two objectives, where one of the objectives dominates the other objective. In the proposed problem, the learner obtains a random reward vector, where each component of the reward vector corresponds to one of the objectives and the distribution of the reward depends on the context that is provided to the learner at the beginning of each round. We call this problem contextual multi-armed bandit with a dominant objective (CMAB-DO). In CMAB-DO, the goal of the learner is to maximize its total reward in the non-dominant objective while ensuring that it maximizes its total reward in the dominant objective. In this case, the optimal arm given the 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. First, we show that the optimal arm lies in the Pareto front. Then, we propose the multi-objective contextual multi-armed bandit algorithm (MOC-MAB), and define two performance measures: the 2-dimensional (2D) regret and the Pareto regret. We show that both the 2D regret and the Pareto regret of MOC-MAB are sublinear in the number of rounds. We also compare the performance of the proposed algorithm with other state-of-the-art methods in synthetic and real-world datasets. The proposed 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 MAB, multi-objective MAB, dominant objective, multi-dimensional regret, Pareto regret.**

I. 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, including recommender systems [2], [3], medical screening [4], cognitive radio networks [5], [6] and wireless network monitoring [7] may involve multiple and possibly conflicting objectives. In this work, we propose a multi-objective contextual MAB problem with dominant and non-dominant objectives. For this problem, we construct a multi-objective contextual MAB algorithm named MOC-MAB, which maximizes the 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 in the beginning of each round. Then, it selects one of the available arms and receives a random reward vector, 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 for a given context is defined as the one that maximizes the expected reward of the non-dominant objective among all arms that maximize the expected reward of the dominant objective.

The learner's performance is measured in terms of its regret, which measures the loss that the learner accumulates due to not knowing the reward distributions beforehand. We introduce two new notions of regret: the 2D regret and the Pareto regret. The 2D regret is a vector whose ith component corresponds to the difference between the expected total reward of an oracle in objective i that selects the optimal arm for each context and that of the learner by time T. On the other hand, the Pareto regret measures sum of the distances of the arms selected by the learner to the Pareto front. For this, we extend the Pareto regret proposed in [8] to take into account the dependence of the Pareto front on the context.

We prove that MOC-MAB achieves $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ 2D regret, where d is the dimension of the context 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 both objectives. We also show that the optimal arm lies in the Pareto front, and MOC-MAB also achieves $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ Pareto regret. Then, we argue that it is possible to make the Pareto regret of MOC-MAB $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ by adjusting its parameters, such that the Pareto regret becomes order optimal up to a logarithmic factor [9], but this comes at an expense of making the regret in the non-dominant objective of MOC-MAB linear in the number of rounds.

To the best of our knowledge, our work is the first to formulate a contextual multi-objective MAB problem and prove sublinear bounds on the 2D regret and the Pareto regret. Different from the conference version [1], in this paper we (i) consider the

Manuscript received August 18, 2017; revised March 9, 2018; accepted May 12, 2018. Date of publication May 29, 2018; date of current version June 12, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qingjiang Shi. This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grants 116C043 and 116E229. A preliminary version of this work was presented in part at the IEEE 27th International Workshop on Machine Learning for Signal Processing, Tokyo, Japan, 2017. *(Corresponding author: Cem Tekin.)*

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: cemtekin@ee.bilkent.edu.tr; turgay@ee.bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2018.2841822

TABLE I
COMPARISON OF THE REGRET BOUNDS AND ASSUMPTIONS IN OUR WORK WITH THE RELATED WORKS

MAB algorithm	Regret bound	Multi-objective	Contextual	Linear rewards	Similarity assumption
Contextual Zooming [10]	$\tilde{O}(T^{1-1/(2+d_z)})$	No	Yes	No	Yes
Query-Ad-Clustering [9]	$\tilde{O}(T^{1-1/(2+d_c)})$	No	Yes	No	Yes
SupLinUCB [11]	$\tilde{O}(\sqrt{T})$	No	Yes	Yes	No
Pareto-UCB1 [8]	$O(\log(T))$	Yes	No	No	No
Scalarized-UCB1 [8]	$O(\log(T))$	Yes	No	No	No
MOC-MAB (our work)	$\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ (2D and Pareto regrets) $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ (Pareto regret only)	Yes	Yes	No	Yes

Pareto regret in addition to the 2D regret, (ii) connect our notion of optimality with lexicographic optimality, (iii) provide a high probability bound on the 2D regret, (iv) show how MOC-MAB can be extended to deal with periodically changing expected arm rewards, (v) discuss how CMAB-DO can be extended for more than two objectives, (vi) provide numerical results on multichannel communication and display advertising applications. Our results show that MOC-MAB outperforms its competitors, which are not specifically designed to deal with problems involving dominant and non-dominant objectives. Moreover, the journal version includes all the proofs.

The rest of the paper is organized as follows. Related work is given in Section II. Problem formulation, definitions of the 2D regret and the Pareto regret, and possible applications of CMAB-DO are given in Section III. MOC-MAB is introduced in Section IV, and its regrets are analyzed in Section V. How MOC-MAB can be extended to work under dynamically changing reward distributions and how CMAB-DO can be extended to capture more than two objectives are discussed in Section VI. Illustrative results are presented in Section VII, and concluding remarks are provided in Section VIII.

II. Related Work

In the past decade, many variants of the classical MAB have been introduced (see [12] for a comprehensive discussion). Two notable examples are contextual MAB [10], [13], [14] and multi-objective MAB [8]. While these examples have been studied separately in prior works, in this paper we aim to fuse contextual MAB and multi-objective MAB together. Below, we discuss the related work on the classical MAB, contextual MAB and multi-objective MAB. The differences between our work and related works are summarized in Table I.

A. The Classical MAB

The classical MAB involves K arms with unknown reward distributions. The learner sequentially selects arms and observes noisy reward samples from the selected arms. The goal of the learner is to use the knowledge it obtains through these observations to maximize its long-term reward. For this, the learner needs to identify arms with high rewards without wasting too much time on arms with low rewards. In conclusion, it needs to strike the balance between exploration and exploitation.

A thorough technical analysis of the classical MAB is given in [15], where it is shown that $O(\log T)$ regret is achieved asymptotically by index policies that use upper confidence bounds (UCBs) for the rewards. This result is tight in the sense that there is a matching asymptotic lower bound. Later on, it is shown in [16] that it is possible to achieve $O(\log T)$ regret by using index policies constructed using the sample means of the arm rewards. The first finite-time logarithmic regret bound is given in [17]. Strikingly, the algorithm that achieves this bound computes the arm indices using only the information about the current round, the sample mean arm rewards and the number of times each arm is selected. This line of research has been followed by many others, and new algorithms with tighter regret bounds have been proposed [18].

B. The Contextual MAB

In the contextual MAB, different from the classical MAB, the learner observes a context (side information) at the beginning of each round, which gives a hint about the expected arm rewards in that round. The context naturally arises in many practical applications such as social recommender systems [19], medical diagnosis [20] and big data stream mining [21]. Existing work on contextual MAB can be categorized into three based on how the contexts arrive and how they are related to the arm rewards.

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. For this category, no statistical assumptions are made on how the contexts arrive. However, given a particular context, the arm rewards come from a fixed distribution parameterized by the context.

This problem is considered in [9], and the Query-Ad-Clustering algorithm that achieves $O(T^{1-1/(2+d_c)+\epsilon})$ regret for any $\epsilon > 0$ is proposed, where d_c is the covering dimension of the similarity space. In addition, $\Omega(T^{1-1/(2+d_p)-\epsilon})$ lower bound on the regret, where d_p is the packing dimension of the similarity space, is also proposed in this work. The main idea behind Query-Ad-Clustering is to partition the context set into disjoint sets and to estimate the expected arm rewards for each set in the partition separately. A parallel work [10] proposes the contextual zooming algorithm which partitions the similarity space non-uniformly, according to both sampling frequency and rewards obtained from different regions of the similarity space. It is shown that contextual zooming achieves $\tilde{O}(T^{1-1/(2+d_z)})$

regret, where d_z is the zooming dimension of the similarity space, which is an optimistic version of the covering dimension that depends on the size of the set of near-optimal arms.

In this contextual MAB category, reward estimates are accurate as long as the contexts that lie in the same set of the context set partition are similar to each other. However, when dimension of the context is high, the regret bound becomes almost linear. This issue is addressed in [22], where 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.

The second category assumes that the expected reward of an arm is a linear combination of the elements of the context. For this model, LinUCB algorithm is proposed in [2]. A modified version of this algorithm, named SupLinUCB, is studied in [11], and is shown to achieve $\tilde{O}(\sqrt{Td})$ regret, where d is the dimension of the context. Another work [23] considers LinUCB and SupLinUCB with kernel functions and proposes an algorithm with $\tilde{O}(\sqrt{T\tilde{d}})$ regret, where $\tilde{d}$ is the effective dimension of the kernel feature space.

The third category assumes that the contexts and arm rewards are jointly drawn from a fixed but unknown distribution. For this case, the Epoch-Greedy algorithm with $O(T^{2/3})$ regret is proposed in [13], and more efficient learning algorithms with $\tilde{O}(T^{1/2})$ regret are developed in [14] and [24].

Our problem is similar to the problems in the first category in terms of the context arrivals and existence of the similarity information.

C. The Multi-objective MAB

In the multi-objective MAB, the learner receives a multi-dimensional reward in each round. Since the rewards are no longer scalar, the definition of a benchmark to compare the learner against becomes obscure. Existing work on multi-objective MAB can be categorized into two: the Pareto approach and the 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 the expected reward of an arm a^* is greater than the expected reward of another arm a in at least one objective, and the expected reward of the arm a is not greater than the expected reward of the arm a^* in any objective, then the arm a^* dominates the arm a. This approach is proposed in [8], and a learning algorithm called Pareto-UCB1 that achieves $O(\log T)$ Pareto regret is proposed. Essentially, this algorithm computes UCB indices for each objective-arm pair, and then, uses these indices to estimate the Pareto front arm set, after which it selects an arm randomly from the Pareto front set. A modified version of this algorithm where the indices depend on both the estimated mean and the estimated standard deviation is proposed in [25]. Numerous other variants are also considered in prior works, including the Pareto Thompson sampling algorithm in [26] and the Annealing Pareto algorithm in [27].

On the other hand, in the scalarized approach [8], [28], a random weight is assigned to each objective at each round, from which for each arm a weighted sum of the indices of the objectives are calculated. In short, this method turns the multi-objective MAB into a single-objective MAB. For instance, Scalarized UCB1 in [8] achieves $O(S' \log(T/S'))$ scalarized regret where S' is the number of scalarization functions used by the algorithm.

The regret notion used in the Pareto and the scalarized approaches are very different from our 2D regret notion. In the Pareto approach, the regret at round t is defined as the minimum distance that should be added to the expected reward vector of the chosen arm at round t to move the chosen arm to the Pareto front. On the other hand, scalarized regret is the difference between scalarized expected rewards of the optimal arm and the chosen arm. Different from these definitions, which define the regret as a scalar quantity, we define the 2D regret as a two-dimensional vector. Hence, our goal is to minimize a multi-dimensional regret measure conditioned on the fact that we minimize the regret in the dominant objective. We show that by achieving this, we also minimize the Pareto regret.

In addition to the works mentioned above, several other works consider multi-criteria reinforcement learning problems, where the rewards are vector-valued [29], [30].

III. Problem Description

A. System Model

The system operates in a sequence of rounds indexed by $t \in \{1, 2, \ldots\}$. At the beginning of round t, the learner observes a d-dimensional context denoted by x_t. Without loss of generality, we assume that x_t lies in the context set $\mathcal{X} := [0, 1]^d$. After observing x_t the learner selects an arm a_t from a finite set $\mathcal{A}$, and then, observes a two dimensional random reward $\boldsymbol{r}_t = (r_t^1, r_t^2)$ that depends both on x_t and a_t. Here, r_t^1 and r_t^2 denote the rewards in the dominant and the non-dominant objectives, respectively, and are given by $r_t^1 = \mu_{a_t}^1(x_t) + \kappa_t^1$ and $r_t^2 = \mu_{a_t}^2(x_t) + \kappa_t^2$, where $\mu_a^i(x)$, $i \in \{1, 2\}$ denotes the expected reward of arm a in objective i given context x, and the noise process $\{(\kappa_t^1, \kappa_t^2)\}$ is such that the marginal distribution of κ_t^i, $i \in \{1, 2\}$ is conditionally 1-sub-Gaussian,[1] i.e.,

$$\forall \lambda \in \mathbb{R} \ \mathrm{E}[e^{\lambda \kappa_t^i} | \boldsymbol{a}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\kappa}_{1:t-1}^1, \boldsymbol{\kappa}_{1:t-1}^2] \leq \exp(\lambda^2/2)$$

where $\boldsymbol{b}_{1:t} := (b_1, \ldots, b_t)$. The expected reward vector for context-arm pair (x, a) is denoted by $\boldsymbol{\mu}_a(x) := (\mu_a^1(x), \mu_a^2(x))$.

The set of arms that maximize the expected reward for the dominant objective for context x is given as $\mathcal{A}^*(x) := \arg\max_{a \in \mathcal{A}} \mu_a^1(x)$. Let $\mu_*^1(x) := \max_{a \in \mathcal{A}} \mu_a^1(x)$ denote the expected reward of an arm in $\mathcal{A}^*(x)$ in the dominant objective. 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. Let $\mu_*^2(x) := \max_{a \in \mathcal{A}^*(x)} \mu_a^2(x)$ denote the expected reward of an optimal arm in the non-dominant objective. We use $a^*(x)$ to refer to an optimal arm for context x. The notion of optimality

[1] Examples of 1-sub-Gaussian distributions include the Gaussian distribution with zero mean and unit variance, and any distribution defined over an interval of length 2 with zero mean [31]. Moreover, our results generalize to the case when κ_t^i is conditionally R-sub-Gaussian for $R \geq 1$. This only changes the constant terms that appear in our regret bounds.

that is defined above coincides with lexicographic optimality [32], which is widely used in multicriteria optimization, and has been considered in numerous applications such as achieving fairness in multirate multicast networks [33] and bit allocation for MPEG video coding [34].

We assume that the expected rewards are Hölder continuous in the context, which is a common assumption in the contextual MAB literature [9], [20], [21].

Assumption 1: There exists $L > 0$, $0 < \alpha \leq 1$ such that for all $i \in \{1, 2\}$, $a \in \mathcal{A}$ and $x, x' \in \mathcal{X}$, we have

$$|\mu_a^i(x) - \mu_a^i(x')| \leq L \left\| x - x' \right\|^{\alpha}.$$

Since Hölder continuity implies continuity, for any non-trivial contextual MAB in which the sets of optimal arms in the first objective are different for at least two contexts, there exists at least one context $x \in \mathcal{X}$ for which $\mathcal{A}^*(x)$ is not a singleton. Let $\mathcal{X}^*$ denote the set of contexts for which $\mathcal{A}^*(x)$ is not a singleton. Since we make no assumptions on how contexts arrive, it is possible that majority of contexts that arrive by round T are in set $\mathcal{X}^*$. This implies that contextual MAB algorithms that only aim at maximizing the rewards in the first objective cannot learn the optimal arms for each context.

Another common way to compare arms when the rewards are multi-dimensional is to use the notion of Pareto optimality, which is described below.

Definition 1 (Pareto Optimality): (i) An arm a is *weakly dominated* by arm a' given context x, denoted by $\boldsymbol{\mu}_a(x) \preceq \boldsymbol{\mu}_{a'}(x)$ or $\boldsymbol{\mu}_{a'}(x) \succeq \boldsymbol{\mu}_a(x)$, if $\mu_a^i(x) \leq \mu_{a'}^i(x), \forall i \in \{1, 2\}$.

(ii) An arm a is *dominated* by arm a' given context x, denoted by $\boldsymbol{\mu}_a(x) \prec \boldsymbol{\mu}_{a'}(x)$ or $\boldsymbol{\mu}_{a'}(x) \succ \boldsymbol{\mu}_a(x)$, if it is weakly dominated and $\exists i \in \{1, 2\}$ such that $\mu_a^i(x) < \mu_{a'}^i(x)$.

(iii) Two arms a and a' are incomparable given context x, denoted by $\boldsymbol{\mu}_a(x) || \boldsymbol{\mu}_{a'}(x)$, if neither arm dominates the other.

(iv) An arm is *Pareto optimal* given context x if it is not dominated by any other arm given context x. Given a particular context x, the set of all Pareto optimal arms is called the *Pareto front*, and is denoted by $\mathcal{O}(x)$.

In the following remark, we explain the connection between lexicographic optimality and Pareto optimality.

Remark 1: Note that $a^*(x) \in \mathcal{O}(x)$ for all $x \in \mathcal{X}$ since $a^*(x)$ is not dominated by any other arm. For all $a \in \mathcal{A}$, we have $\mu_*^1(x) \geq \mu_a^1(x)$. By definition of $a^*(x)$ if there exists an arm a for which $\mu_a^2(x) > \mu_*^2(x)$, then we must have $\mu_a^1(x) < \mu_*^1(x)$. Such an arm will be incomparable with $a^*(x)$.

B. Definitions of the 2D Regret and the Pareto Regret

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. Hence, the 2D regret of the learner by round T is defined as the tuple $(\text{Reg}^1(T), \text{Reg}^2(T))$, where

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

for an arbitrary sequence of contexts $x_1, \ldots, x_T$. When $\text{Reg}^1(T) = O(T^{\gamma_1})$ and $\text{Reg}^2(T) = O(T^{\gamma_2})$ we say that the 2D regret is $O(T^{\max(\gamma_1, \gamma_2)})$.

Another interesting performance measure is the Pareto regret [8], which measures the loss of the learner with respect to arms in the Pareto front. To define the Pareto regret, we first define the Pareto suboptimality gap (PSG).

Definition 2 (PSG of an arm): The PSG of an arm $a \in \mathcal{A}$ given context x, denoted by $\Delta_a(x)$, is defined as the minimum scalar $\epsilon \geq 0$ that needs to be added to all entries of $\boldsymbol{\mu}_a(x)$ such that a becomes a member of the Pareto front. Formally,

$$\Delta_a(x) := \inf_{\epsilon \geq 0} \epsilon \text{ s.t. } (\boldsymbol{\mu}_a(x) + \boldsymbol{\epsilon}) \; || \; \boldsymbol{\mu}_{a'}(x), \forall a' \in \mathcal{O}(x)$$

where $\boldsymbol{\epsilon}$ is a 2-dimensional vector, whose entries are ϵ.

Based on the above definition, the Pareto regret of the learner by round T is given by

$$\text{PR}(T) := \sum_{t=1}^{T} \Delta_{a_t}(x_t). \qquad (2)$$

Our goal is to design a learning algorithm whose 2D and Pareto regrets are sublinear functions of T with high probability. This ensures that the average regrets diminish as $T \to \infty$, and hence, enables the learner to perform on par with an oracle that always selects the optimal arms in terms of the average reward.

C. Applications of CMAB-DO

In this subsection we describe four possible applications of CMAB-DO.

1) Multichannel Communication: Consider a multichannel communication application in which a user chooses a channel $Q \in \mathcal{Q}$ and a transmission rate $R \in \mathcal{R}$ in each round after receiving context $x_t := \{\text{SNR}_{Q,t}\}_{Q \in \mathcal{Q}}$, where $\text{SNR}_{Q,t}$ is the transmit signal to noise ratio of channel Q in round t. For instance, if each channel is also allocated to a primary user, then $\text{SNR}_{Q,t}$ can change from round to round due to time varying transmit power constraint in order not to cause outage to the primary user on channel Q.

In this setup, each arm corresponds to a transmission rate-channel pair (R, Q) 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 round t, it receives a 2-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 rate of $a_{R,Q}$ is equal to $\mu_{a_{R,Q}}^2(x_t) = 1 - p_{\text{out}}(R, Q, x_t)$, 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 $a_{R,Q}$, $r_t^1 \in \{0, R/R_{\max}\}$ and $\mu_{a_{R,Q}}^1(x_t) = R(1 - p_{\text{out}}(R, Q, x_t))/R_{\max}$, where $R_{\max}$ is the maximum rate. It is usually the case that the outage probability increases with R, so maximizing the throughput and reliability are usually conflicting objectives.[2] Illustrative results on this application are given in Section VII-B.

2) 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 round 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 and a positive case is correctly identified.

3) Recommender System: Recommender systems involve optimization of multiple metrics like novelty and diversity in addition to accuracy [35], [36]. Below, we describe how a recommender system with accuracy and diversity metrics can be modeled using CMAB-DO.

At the beginning of round t a user with context x_t arrives to the recommender system. Then, an item from set $\mathcal{A}$ is recommended to the user along with a novelty rating box which the user can use to rate the item as novel or not novel.[3] The recommendation is considered to be accurate when the user clicks to the item, and is considered to be novel when the user rates the item as novel.[4] Thus, $r^1_t = 1$ if the user clicks to the item and 0 otherwise. Similarly, $r^2_t = 1$ if the user rates the item as novel and 0 otherwise. The distribution of (r^1_t, r^2_t) depends on x_t and is unknown to the recommender system.

Another closely related application is display advertising [37], where an advertiser can place an ad to the publisher's website for the user currently visiting the website through a payment mechanism. The goal of the advertiser is to maximize its click through rate while keeping the costs incurred through payments at a low level. Thus, it aims at placing an ad only when the current user with context x_t has positive probability of clicking to the ad. Illustrative results on this application are given in Section VII-C.

4) Network Routing: Packet routing in a communication network commonly involves multiple paths. Adaptive packet routing can improve the performance by avoiding congested and faulty links. In many networking problems, it is desirable to minimize energy consumption as well as the delay due to the energy constraints of sensor nodes. For instance, lexicographic optimality is used in [38] to obtain routing flows in a wireless sensor network with energy limited nodes. Moreover, [39] studies a communication network with elastic and inelastic flows, and proposes load-balancing and rate-control algorithms that prioritize satisfying the rate demanded by inelastic traffic.

Algorithm 1: MOC-MAB.

```
Input: $T$, $d$, $L$, $\alpha$, $m$, $\beta$
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 (3)
    Set $a^*_1 = \arg\max_{a \in \mathcal{A}} g^1_{a,p^*}$ (break ties randomly)
    if $u_{a^*_1,p^*} > \beta v$ then
        Select arm $a_t = a^*_1$
    else
        Find set of candidate optimal arms $\hat{\mathcal{A}}^*$ as given in (4)
        Select arm $a_t = \arg\max_{a \in \hat{\mathcal{A}}^*} g^2_{a,p^*}$ (break ties randomly)
    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
```

Given a source destination pair (src, dst) in an energy constrained wireless sensor network, we can formulate routing of the flow from node src to node dst using CMAB-DO. At the beginning of each round, the network manager observes the network state x_t, which can be the normalized round-trip time on some measurement paths. Then, it selects a path from the set of available paths $\mathcal{A}$ and observes the normalized random energy consumption c^1_t and delay c^2_t over the selected path. These costs are converted to rewards by setting $r^1_t = 1 - c^1_t$ and $r^2_t = 1 - c^2_t$.

IV. The Learning Algorithm

We introduce MOC-MAB in this section. Its pseudocode 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 the context was in p and arm a was selected before the current round, (ii) the sample mean of the rewards obtained from rounds prior to the current round in which the context was in p and arm a was selected, i.e., $\hat{\mu}^1_{a,p}$ and $\hat{\mu}^2_{a,p}$ for the dominant and non-dominant objectives, respectively. The idea behind partitioning is to utilize the similarity of arm rewards given in Assumption 1 to learn together for groups of similar contexts. Basically, when the number of sets in the partition is small, the number of past samples that fall into a specific set is large; however, the similarity of the past samples that fall into the same set is small. The optimal partitioning should balance the inaccuracy in arm reward estimates that results form these two conflicting facts.

[2] Note that in this example, given that arm $a_{R,Q}$ is selected, we have $\kappa^1_t = r^1_t - \mu^1_{a_{R,Q}}(x_t)$ and $\kappa^2_t = r^2_t - \mu^2_{a_{R,Q}}(x_t)$. Clearly, both κ^1_t and κ^2_t are zero mean with support in $[-1, 1]$. Hence, they are 1-sub-Gaussian.

[3] An example recommender system that uses this kind of feedback is given in [36].

[4] In reality, it is possible that some users may not provide the novelty rating. These users can be discarded from the calculation of the regret.

At round t, MOC-MAB first identifies the hypercube in $\mathcal{P}$ that contains x_t, which is denoted by p^*.[5] Then, it calculates the following indices for the rewards in the dominant and the non-dominant objectives:

$$g^i_{a,p^*} := \hat{\mu}^i_{a,p^*} + u_{a,p^*},\ i \in \{1, 2\} \tag{3}$$

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 $\hat{\mu}^i_{a,p^*}$.[6] 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 non-vanishing uncertainty term due to context set partitioning. Since this term is non-vanishing, we also name it the *margin of tolerance*. 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.

After calculating the UCBs, 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 that always selects $a_t = a^*_1$ and that randomly chooses one of the arms with the highest index in the dominant objective in case of a tie, both arms will be equally selected in expectation. Hence, due to the noisy rewards, there are sample paths in which arm 2 is selected more than half of the time. For these sample paths, the expected regret in the non-dominant objective is at least $T/2$. 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 $u_{a^*_1,p^*} \le \beta v$, where $\beta > 0$ is a constant. For this, it calculates the set of candidate optimal arms given as

$$\begin{aligned}\hat{\mathcal{A}}^* &:= \left\{a \in \mathcal{A} : g^1_{a,p*} \ge \hat{\mu}^1_{a^*_1,p^*} - u_{a^*_1,p^*} - 2v\right\}\\ &= \left\{a \in \mathcal{A} : \hat{\mu}^1_{a,p*} \ge \hat{\mu}^1_{a^*_1,p^*} - u_{a^*_1,p^*} - u_{a,p^*} - 2v\right\}.\end{aligned} \tag{4}$$

Here, the term $-u_{a^*_1,p^*} - u_{a,p^*} - 2v$ accounts for the joint uncertainty over the sample mean rewards of arms a and a^*_1. Then, MOC-MAB selects $a_t = \arg\max_{a\in\hat{\mathcal{A}}^*} g^2_{a,p^*}$.

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^*.

Remark 2: At each round, finding the set in $\mathcal{P}$ that x_t belongs to requires $O(d)$ computations. Moreover, each of the following processes requires $O(|\mathcal{A}|)$ computations: (i) finding maximum value among the indices of the dominant objective, (ii) creating a candidate set and finding maximum value among the indices of the non-dominant objective. Hence, MOC-MAB requires $O(dT) + O(|\mathcal{A}|T)$ computations in T rounds. In addition, the memory complexity of MOC-MAB is $O(m^d|\mathcal{A}|)$.

Remark 3: MOC-MAB allows the sample mean reward of the selected arm to be less than the sample mean reward of a^*_1 by at most $u_{a^*_1,p^*} + u_{a,p^*} + 2v$. Here, $2v$ term does not vanish as arms get selected since it results from the partitioning of the context set. While setting v based on the time horizon allows the learner to control the regret due to partitioning, in some settings having this non-vanishing term allows MOC-MAB to achieve reward that is much higher than the reward of the oracle in the non-dominant objective. Such an example is given in Section VII-C.

V. Regret Analysis

In this section we prove that both the 2D regret and the Pareto regret of MOC-MAB are sublinear functions of T. Hence, MOC-MAB is average reward optimal in both regrets. First, we introduce the following as preliminaries.

For an event $\mathcal{F}$, let $\mathcal{F}^c$ denote the complement of that event. For all the parameters defined in Section IV, we explicitly use the round index t, when referring to the value of that parameter at the beginning of round t. For instance, $N_{a,p}(t)$ denotes the value of $N_{a,p}$ at the beginning of round t. Let $N_p(t)$ denote the number of context arrivals to $p \in \mathcal{P}$ by the end of round t, $\tau_p(t)$ denote the round in which a context arrives to $p \in \mathcal{P}$ for the tth time, and $R^i_a(t)$ denote the random reward of arm a in objective i in round t. Let $\tilde{x}_p(t) := x_{\tau_p(t)}$, $\tilde{R}^i_{a,p}(t) := R^i_a(\tau_p(t))$, $\tilde{N}_{a,p}(t) := N_{a,p}(\tau_p(t))$, $\tilde{\mu}^i_{a,p}(t) := \hat{\mu}^i_{a,p}(\tau_p(t))$, $\tilde{a}_p(t) := a_{\tau_p(t)}$, $\tilde{\kappa}^i_p(t) := \kappa^i_{\tau_p(t)}$ and $\tilde{u}_{a,p}(t) := u_{a,p}(\tau_p(t))$. Let $\mathcal{T}_p := \{t \in \{1, \dots, T\} : x_t \in p\}$ denote the set of rounds for which the context is in $p \in \mathcal{P}$.

Next, we define the following lower and upper bounds: $L^i_{a,p}(t) := \tilde{\mu}^i_{a,p}(t) - \tilde{u}_{a,p}(t)$ and $U^i_{a,p}(t) := \tilde{\mu}^i_{a,p}(t) + \tilde{u}_{a,p}(t)$ for $i \in \{1, 2\}$. Let

$$\mathrm{UC}^i_{a,p} := \bigcup_{t=1}^{N_p(T)} \{\mu^i_a(\tilde{x}_p(t)) \notin [L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]\}$$

denote the event that the learner is not confident about its reward estimate in objective i for at least once in rounds in which the context is in p by time T. Here $L^i_{a,p}(t) - v$ and $U^i_{a,p}(t) + v$ are the lower confidence bound (LCB) and UCB for $\mu^i_a(\tilde{x}_p(t))$, respectively. Also, let $\mathrm{UC}^i_p := \cup_{a\in\mathcal{A}}\mathrm{UC}^i_{a,p}$, $\mathrm{UC}_p := \cup_{i\in\{1,2\}}\mathrm{UC}^i_p$ and $\mathrm{UC} := \cup_{p\in\mathcal{P}}\mathrm{UC}_p$, and for each $i \in \{1, 2\}$, $p \in \mathcal{P}$ and $a \in \mathcal{A}$, let

$$\overline{\mu}^i_{a,p} = \sup_{x\in p} \mu^i_a(x) \text{ and } \underline{\mu}^i_{a,p} = \inf_{x\in p} \mu^i_a(x).$$

[5] If the context arrives to the boundary of multiple hypercubes, then it is randomly assigned to one of them.

[6] Although MOC-MAB requires T as input, it can run without the knowledge of T beforehand by applying a method called the doubling-trick. See [40] and [20] for a discussion on the doubling-trick.

Let

$$\mathrm{Reg}^i_p(T) := \sum_{t=1}^{N_p(T)} \mu^i_*(\tilde{x}_p(t)) - \sum_{t=1}^{N_p(T)} \mu^i_{\tilde{a}_p(t)}(\tilde{x}_p(t))$$

denote the regret incurred in objective i for rounds in $\mathcal{T}_p$ (regret incurred in $p \in \mathcal{P}$). Then, the total regret in objective i can be written as

$$\mathrm{Reg}^i(T) = \sum_{p \in \mathcal{P}} \mathrm{Reg}^i_p(T). \tag{5}$$

Thus, the expected regret in objective i becomes

$$\mathrm{E}[\mathrm{Reg}^i(T)] = \sum_{p \in \mathcal{P}} \mathrm{E}[\mathrm{Reg}^i_p(T)]. \tag{6}$$

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

$$\begin{aligned}
&\mathrm{E}[\mathrm{Reg}^i_p(T)] \\
&= \mathrm{E}[\mathrm{Reg}^i_p(T)|\mathrm{UC}]\Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i_p(T)|\mathrm{UC}^c]\Pr(\mathrm{UC}^c) \\
&\le C^i_{\max} N_p(T)\Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i_p(T)|\mathrm{UC}^c] \qquad (7)
\end{aligned}$$

where $C^i_{\max}$ is the maximum difference in the expected reward of an optimal arm and any other arm for objective i.

Having obtained the decomposition in (7), we proceed by bounding the terms in (7). For this, we first bound $\Pr(\mathrm{UC}_p)$ in the next lemma.

Lemma 1: For any $p \in \mathcal{P}$, we have $\Pr(\mathrm{UC}_p) \le 1/(m^d T)$.

Proof: The proof is given in Appendix A. ■

Using the result of Lemma 1, we obtain

$$\Pr(\mathrm{UC}) \le 1/T \text{ and } \Pr(\mathrm{UC}^c) \ge 1 - 1/T. \tag{8}$$

To prove the lemma above, we use the concentration inequality given in Lemma 6 in [31] to bound the probability of $\mathrm{UC}^i_{a,p}$. 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 mean values of $\tilde{R}^i_{a,p}(1), \ldots, \tilde{R}^i_{a,p}(N_p(t))$ different. To overcome this problem, we use the sandwich technique proposed in [20] in order to bound the rewards sampled from actual context arrivals between the rewards sampled from two specific processes that are related to the original process, where each process has a fixed mean value.

After bounding the probability of the event $\Pr(\mathrm{UC}_p)$, we bound the instantaneous (single round) regret on event $\Pr(\mathrm{UC}^c)$. For simplicity of notation, in the following lemmas we use $a^*(t) := a^*(\tilde{x}_p(t))$ to denote the optimal arm, $\tilde{a}(t) := \tilde{a}_p(t)$ to denote the arm selected at round $\tau_p(t)$ and $\hat{a}^*_1(t)$ to denote the arm whose first index is highest at round $\tau_p(t)$, when the set $p \in \mathcal{P}$ that the context belongs to is obvious.

The following lemma shows that on event UC^c_p the regret incurred in a round $\tau_p(t)$ for the dominant objective can be bounded as function of the difference between the upper and lower confidence bounds plus the margin of tolerance.

Lemma 2: When MOC-MAB is run, on event UC^c_p, we have

$$\begin{aligned}
\mu^1_{a^*(t)}(\tilde{x}_p(t)) - \mu^1_{\tilde{a}(t)}(\tilde{x}_p(t)) \le U^1_{\tilde{a}(t),p}(t) - L^1_{\tilde{a}(t),p}(t) \\
+ 2(\beta+2)v
\end{aligned}$$

for all $t \in \{1, \ldots, N_p(T)\}$.

Proof: We consider two cases. When $\tilde{u}_{\hat{a}^*_1(t),p}(t) \le \beta v$, we have

$$\begin{aligned}
U^1_{\tilde{a}(t),p}(t) &\ge L^1_{\hat{a}^*_1(t),p}(t) - 2v \\
&\ge U^1_{\hat{a}^*_1(t),p}(t) - 2\tilde{u}_{\hat{a}^*_1(t),p}(t) - 2v \\
&\ge U^1_{\hat{a}^*_1(t),p}(t) - 2(\beta+1)v.
\end{aligned}$$

On the other hand, when $\tilde{u}_{\hat{a}^*_1(t),p}(t) > \beta v$, the selected arm is $\tilde{a}(t) = \hat{a}^*_1(t)$. Hence, we obtain

$$U^1_{\tilde{a}(t),p}(t) = U^1_{\hat{a}^*_1(t),p}(t) \ge U^1_{\hat{a}^*_1(t),p}(t) - 2(\beta+1)v.$$

Thus, for both cases, we have

$$U^1_{\tilde{a}(t),p}(t) \ge U^1_{\hat{a}^*_1(t),p}(t) - 2(\beta+1)v \tag{9}$$

and

$$U^1_{\hat{a}^*_1(t),p}(t) \ge U^1_{a^*(t),p}(t). \tag{10}$$

On event UC^c_p, we also have

$$\mu^1_{a^*(t)}(\tilde{x}_p(t)) \le U^1_{a^*(t),p}(t) + v \tag{11}$$

and

$$\mu^1_{\tilde{a}(t)}(\tilde{x}_p(t)) \ge L^1_{\tilde{a}(t),p}(t) - v. \tag{12}$$

By combining (9)–(12), we obtain

$$\begin{aligned}
\mu^1_{a^*(t)}(\tilde{x}_p(t)) - \mu^1_{\tilde{a}(t)}(\tilde{x}_p(t)) \le U^1_{\tilde{a}(t),p}(t) - L^1_{\tilde{a}(t),p}(t) \\
+ 2(\beta+2)v.
\end{aligned}$$

■

The lemma below bounds the regret incurred in a round $\tau_p(t)$ for the non-dominant objective on event UC^c_p when the uncertainty level of the arm with the highest index in the dominant objective is low.

Lemma 3: When MOC-MAB is run, on event UC^c_p, for $t \in \{1, \ldots, N_p(T)\}$ if

$$\tilde{u}_{\hat{a}^*_1(t),p}(t) \le \beta v$$

holds, then we have

$$\mu^2_{a^*(t)}(\tilde{x}_p(t)) - \mu^2_{\tilde{a}(t)}(\tilde{x}_p(t)) \le U^2_{\tilde{a}(t),p}(t) - L^2_{\tilde{a}(t),p} + 2v.$$

Proof: When $\tilde{u}_{\hat{a}^*_1(t),p}(t) \le \beta v$ holds, all arms that are selected as candidate optimal arms have their index for objective 1 in the interval $[L^1_{\hat{a}^*_1(t),p}(t) - 2v, U^1_{\hat{a}^*_1(t),p}(t)]$. Next, we show that $U^1_{a^*(t),p}(t)$ is also in this interval.

On event UC^c_p, we have

$$\begin{aligned}
\mu^1_{a^*(t)}(\tilde{x}_p(t)) &\in [L^1_{a^*(t),p}(t) - v, U^1_{a^*(t),p}(t) + v] \\
\mu^1_{\hat{a}^*_1(t)}(\tilde{x}_p(t)) &\in [L^1_{\hat{a}^*_1(t),p}(t) - v, U^1_{\hat{a}^*_1(t),p}(t) + v].
\end{aligned}$$

We also know that

$$\mu^1_{a^*(t)}(\tilde{x}_p(t)) \geq \mu^1_{\hat{a}_1^*(t)}(\tilde{x}_p(t)).$$

Using the inequalities above, we obtain

$$U^1_{a^*(t),p}(t) \geq \mu^1_{a^*(t)}(\tilde{x}_p(t)) - v \geq \mu^1_{\hat{a}_1^*(t)}(\tilde{x}_p(t)) - v \\ \geq L^1_{\hat{a}_1^*(t),p}(t) - 2v.$$

Since the selected arm has the maximum index for the non-dominant objective among all arms whose indices for the dominant objective are in $[L^1_{\hat{a}_1^*(t),p}(t) - 2v, U^1_{\hat{a}_1^*(t),p}(t)]$, we have $U^2_{\tilde{a}(t),p}(t) \geq U^2_{a^*(t),p}(t)$. Combining this with the fact that UC^c_p holds, we get

$$\mu^2_{\tilde{a}(t)}(\tilde{x}_p(t)) \geq L^2_{\tilde{a}(t),p}(t) - v \tag{13}$$

and

$$\mu^2_{a^*(t)}(\tilde{x}_p(t)) \leq U^2_{a^*(t),p}(t) + v \leq U^2_{\tilde{a}(t),p}(t) + v. \tag{14}$$

Finally, by combining (13) and (14), we obtain

$$\mu^2_{a^*(t)}(\tilde{x}_p(t)) - \mu^2_{\tilde{a}(t)}(\tilde{x}_p(t)) \leq U^2_{\tilde{a}(t),p}(t) - L^2_{\tilde{a}(t),p}(t) + 2v.$$

■

For any $p \in \mathcal{P}$, we also need to bound the regret of the non-dominant objective for rounds in which $\tilde{u}_{\hat{a}_1^*(t),p}(t) > \beta v$, $t \in \{1, \ldots, N_p(T)\}$.

Lemma 4: When MOC-MAB is run, the number of rounds in $\mathcal{T}_p$ for which $\tilde{u}_{\hat{a}_1^*(t),p}(t) > \beta v$ happens is bounded above by

$$|\mathcal{A}| \left(\frac{2A_{m,T}}{\beta^2 v^2} + 1 \right).$$

Proof: This event happens when $\tilde{N}_{\hat{a}_1^*(t),p}(t) < 2A_{m,T}/(\beta^2 v^2)$. Every such event will result in an increase in the value of $N_{\hat{a}_1^*(t),p}$ by one. Hence, for $p \in \mathcal{P}$ and $a \in \mathcal{A}$, the number of times $\tilde{u}_{a,p}(t) > \beta v$ can happen is bounded above by $2A_{m,T}/(\beta^2 v^2) + 1$. The final result is obtained by summing over all arms. ■

In the next lemmas, we bound $\mathrm{Reg}^1_p(t)$ and $\mathrm{Reg}^2_p(t)$ given that UC^c holds.

Lemma 5: When MOC-MAB is run, on event UC^c, we have for all $p \in \mathcal{P}$

$$\mathrm{Reg}^1_p(t) \leq |\mathcal{A}| C^1_{\max} + 2B_{m,T}\sqrt{|\mathcal{A}| N_p(t)} + 2(\beta + 2) v N_p(t).$$

where $B_{m,T} := 2\sqrt{2A_{m,T}}$.

Proof: The proof is given in Appendix B. ■

Lemma 6: When MOC-MAB is run, on event UC^c we have for all $p \in \mathcal{P}$

$$\mathrm{Reg}^2_p(t) \leq C^2_{\max} |\mathcal{A}| \left(\frac{2A_{m,T}}{\beta^2 v^2} + 1 \right) + 2vN_p(t) \\ + 2B_{m,T}\sqrt{|\mathcal{A}| N_p(t)}.$$

Proof: The proof is given in Appendix C. ■

Next, we use the result of Lemmas 1, 5 and 6 to find a bound on $\mathrm{Reg}^i(t)$ that holds for all $t \leq T$ with probability at least $1 - 1/T$.

Theorem 1: When MOC-MAB is run, we have for any $i \in \{1, 2\}$

$$\Pr(\mathrm{Reg}^i(t) < \epsilon_i(t)\, \forall t \in \{1, \ldots, T\}) \geq 1 - 1/T$$

where

$$\epsilon_1(t) = m^d |\mathcal{A}| C^1_{\max} + 2B_{m,T}\sqrt{|\mathcal{A}| m^d t} + 2(\beta + 2) v t$$

and

$$\epsilon_2(t) = m^d |\mathcal{A}| C^2_{\max} + m^d C^2_{\max} |\mathcal{A}| \left(\frac{2A_{m,T}}{\beta^2 v^2} \right) \\ + 2B_{m,T}\sqrt{|\mathcal{A}| m^d t} + 2vt.$$

Proof: By (5) and Lemmas 5 and 6, we have on event UC^c:

$$\mathrm{Reg}^1(t) \leq m^d |\mathcal{A}| C^1_{\max} + 2B_{m,T} \sum_{p \in \mathcal{P}} \sqrt{|\mathcal{A}| N_p(t)} \\ + 2(\beta + 2) v t \\ \leq m^d |\mathcal{A}| C^1_{\max} + 2B_{m,T}\sqrt{|\mathcal{A}| m^d t} \\ + 2(\beta + 2) v t$$

and

$$\mathrm{Reg}^2(t) \leq m^d |\mathcal{A}| C^2_{\max} + m^d C^2_{\max} |\mathcal{A}| \left(\frac{2A_{m,T}}{\beta^2 v^2} \right) \\ + 2B_{m,T} \sum_{p \in \mathcal{P}} \sqrt{|\mathcal{A}| N_p(t)} + 2vt \\ \leq m^d |\mathcal{A}| C^2_{\max} + m^d C^2_{\max} |\mathcal{A}| \left(\frac{2A_{m,T}}{\beta^2 v^2} \right) \\ + 2B_{m,T}\sqrt{|\mathcal{A}| m^d t} + 2vt$$

for all $t \leq T$. The result follows from the fact that UC^c holds with probability at least $1 - 1/T$. ■

The following theorem shows that the expected 2D regret of MOC-MAB by time T is $\tilde{O}(T^{\frac{2\alpha+d}{3\alpha+d}})$.

Theorem 2: When MOC-MAB is run with inputs $m = \lceil T^{1/(3\alpha+d)} \rceil$ and $\beta > 0$, we have

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

and

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

Proof: $\mathrm{E}[\mathrm{Reg}^i(T)]$ is bounded by using the result of Theorem 1 and (7):

$$\begin{aligned}\mathrm{E}[\mathrm{Reg}^i(T)] &\le \mathrm{E}[\mathrm{Reg}^i(T)|\mathrm{UC}^c] + \sum_{p\in\mathcal{P}} C^i_{\max} N_p(T)\Pr(\mathrm{UC})\\ &\le \mathrm{E}[\mathrm{Reg}^i(T)|\mathrm{UC}^c] + \sum_{p\in\mathcal{P}} C^i_{\max} N_p(T)/T\\ &= \mathrm{E}[\mathrm{Reg}^i(T)|\mathrm{UC}^c] + C^i_{\max}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathrm{E}[\mathrm{Reg}^1(T)] &\le \epsilon_1(T) + C^1_{\max}\\ \mathrm{E}[\mathrm{Reg}^2(T)] &\le \epsilon_2(T) + C^2_{\max}.\end{aligned}$$

It can be shown that when we set $m = \lceil T^{1/(2\alpha+d)}\rceil$ regret bound of the dominant objective becomes $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ and regret bound of the non-dominant objective becomes $O(T)$. The optimal value for m that makes both regrets sublinear is $m = \lceil T^{1/(3\alpha+d)}\rceil$. With this value of m, we obtain

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

and

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

■

From the results above we conclude that both regrets are $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$, where for the first regret bound the constant that multiplies the highest order of the regret does not depend on $\mathcal{A}$, while the dependence on this term is linear for the second regret bound.

Next, we show that the expected value of the Pareto regret of MOC-MAB given in (2) is also $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$.

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

$$\Pr(\mathrm{PR}(t) < \epsilon_1(t)\,\forall t \in \{1,\dots,T\}) \ge 1 - 1/T$$

where $\epsilon_1(t)$ is given in Theorem 1 and

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

Proof: Consider any $p \in \mathcal{P}$ and $t \in \{1,\dots,N_p(T)\}$. By definition $\Delta_{\tilde{a}(t)}(\tilde{x}_p(t)) \le \mu^1_{a^*(t)}(\tilde{x}_p(t)) - \mu^1_{\tilde{a}(t)}(\tilde{x}_p(t))$. This holds since for any $\epsilon > 0$, adding $\mu^1_{a^*(t)}(\tilde{x}_p(t)) - \mu^1_{\tilde{a}(t)}(\tilde{x}_p(t)) + \epsilon$ to $\mu^1_{\tilde{a}(t)}(\tilde{x}_p(t))$ will either make it (i) dominate the arms in $\mathcal{O}(\tilde{x}_p(t))$ or (ii) incomparable with the arms in $\mathcal{O}(\tilde{x}_p(t))$. Hence, using the result in Lemma 2, we have on event UC^c

$$\Delta_{\tilde{a}(t)}(\tilde{x}_p(t)) \le U^1_{\tilde{a}(t),p}(t) - L^1_{\tilde{a}(t),p}(t) + 2(\beta+2)v.$$

Let $\mathrm{PR}_p(T) := \sum_{t=1}^{N_p(T)} \Delta_{\tilde{a}(t)}(\tilde{x}_p(t))$. Hence, $\mathrm{PR}(T) = \sum_{p\in\mathcal{P}} \mathrm{PR}_p(T)$. Due to this, the results derived for $\mathrm{Reg}^1(t)$ and $\mathrm{Reg}^1(T)$ in Theorems 1 and 2 also hold for $\mathrm{PR}_p(t)$ and $\mathrm{PR}_p(T)$. ■

Theorems 2 and 3 show that the regret measures $\mathrm{E}[\mathrm{Reg}^1(T)]$, $\mathrm{E}[\mathrm{Reg}^2(T)]$ and $\mathrm{E}[\mathrm{PR}(T)]$ for MOC-MAB are all $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ when it is run with $m = \lceil T^{1/(3\alpha+d)}\rceil$. This implies that MOC-MAB is average reward optimal in all regret measures as $T \to \infty$. The growth rate of the Pareto regret can be further decreased by setting $m = \lceil T^{1/(2\alpha+d)}\rceil$. This will make the Pareto regret $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ (which matches with the lower bound in [9] for the single-objective contextual MAB with similarity information up to a logaritmic factor) but will also make the regret in the non-dominant objective linear.

VI. Extensions

A. Learning Under Periodically Changing Reward Distributions

In many practical cases, the reward distribution of an arm changes periodically over time even under the same context. For instance, in a recommender system the probability that a user clicks to an ad may change with the time of the day, but the pattern of change can be periodical on a daily basis and this can be known by the system. Moreover, this change is usually gradual over time. In this section, we extend MOC-MAB such that it can deal with such settings.

For this, let T_s denote the period. For the d-dimensional context $x_t = (x_{1,t}, x_{2,t}, ..., x_{d,t})$ received at round t let $\hat{x}_t := (x_{1,t}, x_{2,t}, ..., x_{d+1,t})$ denote the extended context where $x_{d+1,t} := (t \mod T_s)/T_s$ is the time context. Let $\hat{\mathcal{X}}$ denote the $d+1$ dimensional extended context set constructed by adding the time dimension to $\mathcal{X}$. It is assumed that the following holds for the extended contexts.

Assumption 2: Given any $\hat{x}, \hat{x}' \in \hat{\mathcal{X}}$, there exists $\hat{L} > 0$ and $0 < \hat{\alpha} \le 1$ such that for all $i \in \{1, 2\}$ and $a \in \mathcal{A}$, we have

$$|\mu^i_a(\hat{x}) - \mu^i_a(\hat{x}')| \le \hat{L}||\hat{x} - \hat{x}'||^{\hat{\alpha}}.$$

Note that Assumption 2 implies Assumption 1 with $L = \hat{L}$ and $\alpha = \hat{\alpha}$ when $\hat{x}_{d+1} = \hat{x}'_{d+1}$. Moreover, for two contexts $(x_1, \dots, x_d, x_{d+1})$ and $(x_1, \dots, x_d, x'_{d+1})$, we have

$$|\mu^i_a(\hat{x}) - \mu^i_a(\hat{x}')| \le \hat{L}|x_{d+1} - x'_{d+1}|^{\hat{\alpha}}$$

which implies that the change in the expected rewards is gradual. Under Assumption 2, the performance of MOC-MAB is bounded as follows.

Corollary 1: When MOC-MAB is run with inputs $\hat{L}, \hat{\alpha}, m = \lceil T^{1/(3\hat{\alpha}+d+1)}\rceil$, and $\beta > 0$ by using the extended context set $\hat{\mathcal{X}}$ instead of the original context set $\mathcal{X}$, we have

$$\mathrm{E}[\mathrm{Reg}^i(T)] = \tilde{O}(T^{(2\hat{\alpha}+d+1)/(3\hat{\alpha}+d+1)}) \text{ for } i \in \{1, 2\}.$$

Proof: The proof simply follows from the proof of Theorem 2 by extending the dimension of the context set by one. ■

B. Lexicographic Optimality for $d_r > 2$ Objectives

Our problem formulation can be generalized to handle $d_r > 2$ objectives as follows. Let $\boldsymbol{r}_t := (r_t^1, \ldots, r_t^{d_r})$ denote the reward vector in round t and $\boldsymbol{\mu}_a(x) := (\mu_a^1(x), \ldots, \mu_a^{d_r}(x))$ denote the expected reward vector for context-arm pair (x, a). We say that arm a lexicographically dominates arm a' in the first j objectives for context x, denoted by $\boldsymbol{\mu}_a(x) >_{\text{lex},j} \boldsymbol{\mu}_{a'}(x)$ if $\mu_a^i(x) > \mu_{a'}^i(x)$, where $i := \min\{k \leq j : \mu_a^k(x) \neq \mu_{a'}^k(x)\}$.[7] Then, arm a is defined to be lexicographically optimal for context x if there is no other arm that lexicographically dominates it in d_r objectives.

Let $\mu_*^i(x)$ denote the expected reward of a lexicographically optimal arm for context x in objective i. Then, the d_r-dimensional regret is defined as follows:

$$\mathbf{Reg}(T) := (\text{Reg}^1(T), \ldots, \text{Reg}^{d_r}(T)) \text{ where}$$

$$\text{Reg}^i(T) := \sum_{t=1}^{T} \mu_*^i(x_t) - \sum_{t=1}^{T} \mu_{a_t}^i(x_t), i \in \{1, \ldots, d_r\}.$$

Generalizing MOC-MAB to achieve sublinear regret for all objectives will require construction of a hierarchy of candidate optimal arm sets similar to the one given in (4). We leave this interesting research problem as future work, and explain when lexicographically optimality in the first two objectives indicates lexicographic optimality in d_r objectives and why the number of cases in which lexicographically optimality in the first two objectives does not indicate lexicographic optimality in d_r objectives is *scarce*.

Let $\mathcal{A}_j^*(x)$ denote the set of lexicographically optimal arms for context x in the first j objectives. We call the case $\mathcal{A}_2^*(x) = \mathcal{A}_{d_r}^*(x)$ for all $x \in \mathcal{X}$ the *degenerate case* of the d_r-objective contextual MAB. Similarly, we call the case when there exists some $x \in \mathcal{X}$, for which $\mathcal{A}_2^*(x) \neq \mathcal{A}_{d_r}^*(x)$ as the *non-degenerate case* of the d_r-objective contextual MAB. Next, we argue that the non-degenerate case is uncommon. Since $\mathcal{A}_j^*(x) \supseteq \mathcal{A}_{j+1}^*(x)$ for $j \in \{1, \ldots, d_r - 1\}$ and there is at least one lexicographically optimal arm, $\mathcal{A}_2^*(x) \neq \mathcal{A}_{d_r}^*(x)$ implies that $\mathcal{A}_2^*(x)$ is not a singleton. This implies existence of two arms a and b such that $\mu_a^1(x) = \mu_b^1(x)$ and $\mu_a^2(x) = \mu_b^2(x)$. In contrast, for the contextual MAB to be non-trivial, we only require existence of at least one context $x \in \mathcal{X}$ and arms a and b such that $\mu_a^1(x) = \mu_b^1(x)$.

VII. Illustrative Results

In order to evaluate the performance of MOC-MAB, we run three different experiments both with synthetic and real-world datasets.

We compare MOC-MAB with the following MAB algorithms:

Pareto UCB1 (P-UCB1): This is the Empirical Pareto UCB1 algorithm proposed in [8].

Scalarized UCB1 (S-UCB1): This is the Scalarized Multi-objective UCB1 algorithm proposed in [8].

Contextual Pareto UCB1 (CP-UCB1): This is the contextual version of P-UCB1 which partitions the context set 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 set 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 [17], which partitions the context set 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.

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 set is formed by choosing m according to Theorem 2, and L and α are taken as 1. For MOC-MAB, β is chosen as 1 unless stated otherwise. In addition, we scaled down the uncertainty level (also known as the confidence term or the inflation term) of all the algorithms by a constant chosen from $\{1, 1/5, 1/10, 1/15, 1/20, 1/25, 1/30\}$, since we observed that the regrets of the algorithms in the dominant objective may become smaller when the uncertainty level is scaled down. The reported results correspond to runs performed using the optimal scale factor for each experiment.

A. Experiment 1 - Synthetic Dataset

In this experiment, we compare MOC-MAB with other MAB algorithms on a synthetic multi-objective dataset. We take $\mathcal{X} = [0, 1]^2$ and assume that the context at each round is chosen uniformly at random from $\mathcal{X}$. We consider 4 arms and the time horizon is set as $T = 10^5$. The expected arm rewards for 3 of the arms are generated as follows: We generate 3 multivariate Gaussian distributions for the dominant objective and 3 multivariate Gaussian distributions for the non-dominant objective. For the dominant objective, the mean vectors of the first two distributions are set as $[0.3, 0.5]$, and the mean vector of the third distribution is set as $[0.7, 0.5]$. Similarly, for the non-dominant objective, the mean vectors of the distributions are set as $[0.3, 0.7]$, $[0.3, 0.3]$ and $[0.7, 0.5]$, respectively. For all the Gaussian 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 it with a constant, such that its maximum value becomes 1. These normalized distributions form the expected arm rewards. In addition, the expected reward of the fourth arm for the dominant objective is set as 0, and its expected reward for the non-dominant objective is set as the normalized multivariate Gaussian distribution with mean vector $[0.7, 0.5]$. We assume that the 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 distribution at context x.

Every algorithm is run 100 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

[7] If i does not exist then $\mu_a^k(x) = \mu_{a'}^k(x)$ for all $k \in \{1, \ldots, j\}$, and hence, arm a does not lexicographically dominate arm a' in the first j objectives.

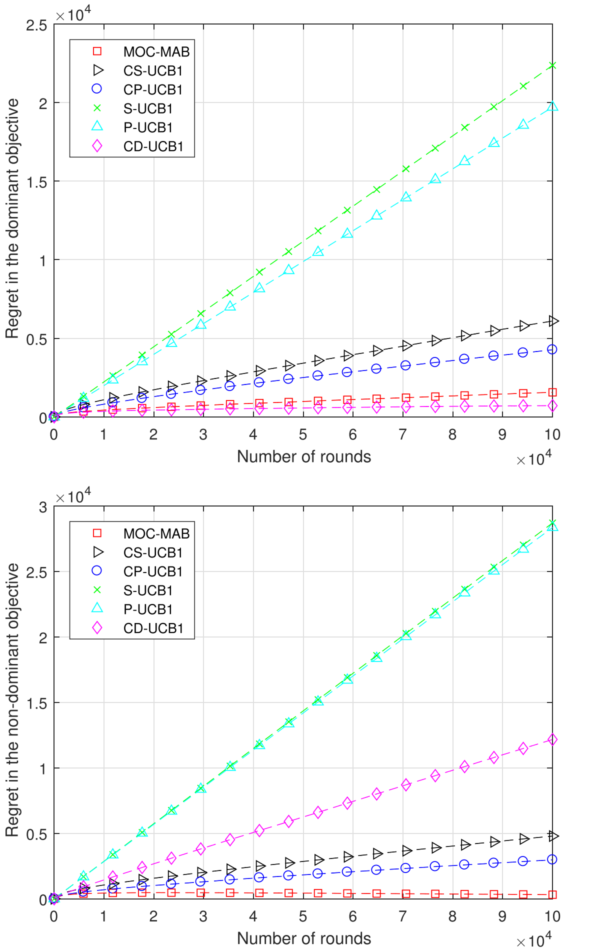

Fig. 1. Regrets of MOC-MAB and the other algorithms for Experiment 1.

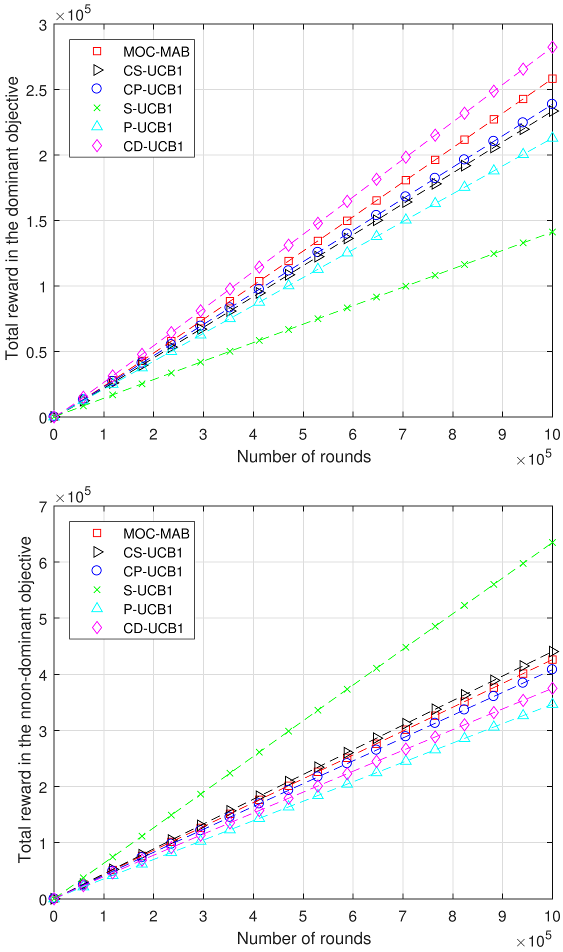

Fig. 2. Total rewards of MOC-MAB and the other algorithms for Experiment 2.

function of time (rounds). As observed from the results, MOC-MAB beats all other algorithms in both objectives except CD-UCB1. While the regret of CD-UCB1 in the dominant objective is slightly better than that of MOC-MAB, its regret is much worse than MOC-MAB in the non-dominant objective. This is expected since it only aims to maximize the reward in the dominant objective without considering the other objective.

B. Experiment 2 - Multichannel Communication

In this experiment, we consider the multichannel communication application given in Section III-C with $\mathcal{Q} = \{1, 2\}$, $\mathcal{R} = \{1, 0.5, 0.25, 0.1\}$ and $T = 10^6$. The channel gain for channel Q in round t, denoted by $h^2_{Q,t}$ is independently sampled from the exponential distribution with parameter λ_Q, where $[\lambda_1, \lambda_2] = [0.25\, 0.25]$. The type of the distributions and the parameters are unknown to the user. $\text{SNR}_{Q,t}$ is sampled from the uniform distribution over $[0, 5]$ independently for both channels. In this case, the outage event for transmission rate-channel pair (R, Q) in round t is defined as $\log_2(1 + h^2_{Q,t}\text{SNR}_{Q,t}) < R$.

Every algorithm is run 20 times and the results are averaged over these runs. Simulation results given in Fig. 2 show the total reward of the algorithms in both objectives as a function of rounds. As observed from the results, there is no algorithm that beats MOC-MAB in both objectives. In the dominant objective, the total reward of MOC-MAB is 8.21% higher than that of CP-UCB1, 10.59% higher than that of CS-UCB1, 21.33% higher than that of P-UCB1 and 82.94% higher than that of S-UCB1 but 8.52% lower than that of CD-UCB1. Similar to Experiment 1, we expect the total reward of CD-UCB1 to be higher than MOC-MAB because it neglects the non-dominant objective. On the other hand, in the non-dominant objective, MOC-MAB achieves total reward 13.66% higher than that of CD-UCB1.

C. Experiment 3 - Display Advertising

In this experiment, we consider a simplified display advertising model where in each round t a user with context x^{usr}_t visits a publisher's website, an ad with context x^{ad}_t arrives to an advertiser, which together constitute the context $x_t = (x^{\text{usr}}_t, x^{\text{ad}}_t)$. Then, the advertiser decides whether to display the ad on the publisher's website (indicated by action a) or not (indicated by action b). The advertiser makes a unit payment to the publisher for each displayed ad (pay-per-view model). The first objective is related to the click through rate and the second objective is related to the average payment. Essentially, when action a is taken in round t, then $r^2_t = 0$, and $r^1_t = 0$ if the user does not

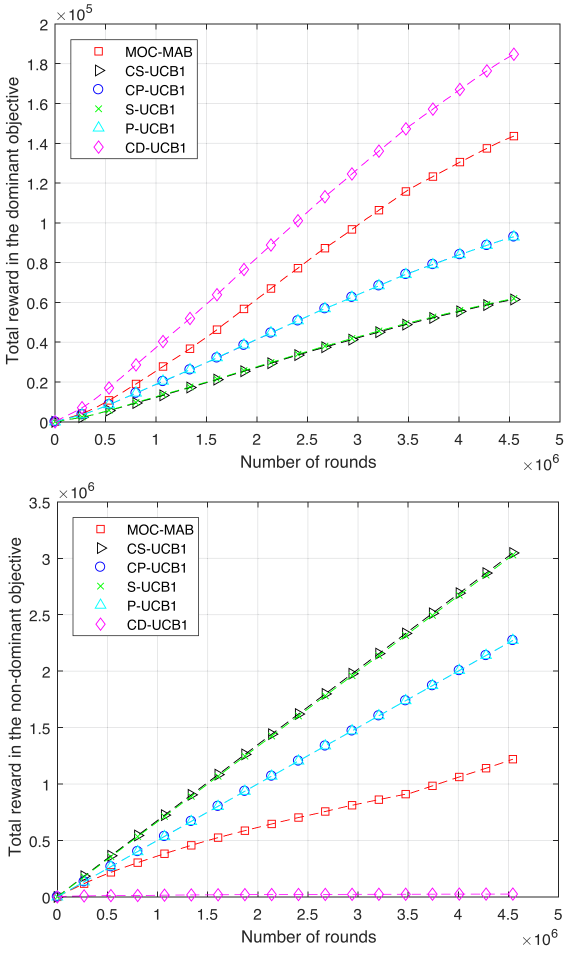

Fig. 3. Total rewards of MOC-MAB and the other algorithms for Experiment 3.

click to the ad and $r_t^1 = 1$ otherwise. When action b is taken in round t, the reward is always $(r_t^1, r_t^2) = (0, 1)$.

We simulate the model described above by using the Yahoo! Webscope dataset R6A,[8] which consists of over 45 million visits to the Yahoo! Today module during 10 days. This dataset was collected from a personalized news recommender system where articles were displayed to users with a picture, title and a short summary, and the click events were recorded. In essence, the dataset only contains a set of continuous features derived from users and news articles by using conjoint analysis and the click events [41]. Thus, for our illustrative result, we adopt the feature of the news article as the feature of the ad and the click event as the event that the user clicks to the displayed ad.

We consider the data collected in the first day which consists of around 4.5 million samples. Each user and item is represented by 6 features, one of which is always 1. We discard the constant features and apply PCA to produce two-dimensional user and item contexts. PCA is applied over all user features to obtain the two-dimensional user contexts x_t^{usr}. To obtain the add contexts x_t^{ad}, we first identify the number of ads with unique features, and then, apply PCA over these. The total number of clicks on day 1 is only 4.07% of the total number of user-ad pairs. Since the click events are scarce, the difference between the empirical rewards of actions a and b in the dominant objective is very small. Thus, we set $\beta = 0.1$ in MOC-MAB in order to further decrease uncertainty in the first objective.

Simulation results given in Fig. 3 show the total reward of the algorithms in both objectives as a function of rounds. In the dominant objective, the total reward of MOC-MAB is 54.5% higher than that of CP-UCB1, 133.6% higher than that of CS-UCB1, 54.5% higher than that of P-UCB1 and 131.8% higher than that of S-UCB1 but 22.3% lower than that of CD-UCB1. In the non-dominant objective, the total reward of MOC-MAB is 46.3% lower than that of CP-UCB1, 60% lower than that of CS-UCB1, 46.3% lower than that of P-UCB1, 59.7% lower than that of S-UCB1 and 4751.9% higher than that of CD-UCB1. As seen from these results, there is no algoritm that outperforms MOC-MAB in both objectives. Although CD-UCB1 outperforms MOC-MAB in the first objective, its total reward in the second objective is much less than the total reward of MOC-MAB.

VIII. Conclusion

In this paper, we propose a new contextual MAB problem with two objectives in which one objective is dominant and the other is non-dominant. According to this definition, we propose two performance metrics: the 2D regret (which is multi-dimensional) and the Pareto regret (which is scalar). Then, we propose an online learning algorithm called MOC-MAB and show that it achieves sublinear 2D regret and Pareto regret. To the best of our knowledge, our work is the first to consider a multi-objective contextual MAB problem where the expected arm rewards and contexts are related through similarity information. We also evaluate the performance of MOC-MAB on both synthetic and real-world datasets and compare it with offline methods and other MAB algorithms. Our results demonstrate that MOC-MAB outperforms its competitors, which are not specifically designed to deal with problems involving dominant and non-dominant objectives.

Appendix A
Proof of Lemma 1

From the definitions of $L^i_{a,p}(t)$, $U^i_{a,p}(t)$ and $\text{UC}^i_{a,p}$, it can be observed that the event $\text{UC}^i_{a,p}$ happens when $\mu^i_a(\tilde{x}_p(t))$ does not fall into the confidence interval $[L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]$ for some t. The probability of this event could be easily bounded by using the concentration inequality given in Appendix D, if the expected reward from the same arm did not change over rounds. However, this is not the case in our model since the elements of $\{\tilde{x}_p(t)\}_{t=1}^{N_p(T)}$ are not identical which makes the distributions of $\tilde{R}^i_{a,p}(t)$, $t \in \{1, \ldots, N_p(T)\}$ different.

In order to resolve this issue, we propose the following: Recall that

$$\tilde{R}^i_{a,p}(t) = \mu^i_a(\tilde{x}_p(t)) + \tilde{\kappa}^i_p(t)$$

[8] http://webscope.sandbox.yahoo.com/

and

$$\tilde{\mu}^i_{a,p}(t) = \frac{\sum_{l=1}^{t-1} \tilde{R}^i_{a,p}(l)\mathrm{I}(\tilde{a}_p(l)=a)}{\tilde{N}_{a,p}(t)}$$

when $\tilde{N}_{a,p}(t) > 0$. Note that when $\tilde{N}_{a,p}(t) = 0$, we have $\tilde{\mu}^i_{a,p}(t) = 0$. We define two new sequences of random variables, whose sample mean values will lower and upper bound $\tilde{\mu}^i_{a,p}(t)$. The *best sequence* is defined as $\{\overline{R}^i_{a,p}(t)\}_{t=1}^{N_p(T)}$ where

$$\overline{R}^i_{a,p}(t) = \overline{\mu}^i_{a,p} + \tilde{\kappa}^i_p(t)$$

and the *worst sequence* is defined as $\{\underline{R}^i_{a,p}(t)\}_{t=1}^{N_p(T)}$ where

$$\underline{R}^i_{a,p}(t) = \underline{\mu}^i_{a,p} + \tilde{\kappa}^i_p(t).$$

Let

$$\overline{\mu}^i_{a,p}(t) := \sum_{l=1}^{t-1} \overline{R}^i_{a,p}(l)\mathrm{I}(\tilde{a}_p(l)=a)/\tilde{N}_{a,p}(t)$$

$$\underline{\mu}^i_{a,p}(t) := \sum_{l=1}^{t-1} \underline{R}^i_{a,p}(l)\mathrm{I}(\tilde{a}_p(l)=a)/\tilde{N}_{a,p}(t)$$

for $\tilde{N}_{a,p}(t) > 0$ and $\overline{\mu}^i_{a,p}(t) = \underline{\mu}^i_{a,p}(t) = 0$ for $\tilde{N}_{a,p}(t) = 0$. We have

$$\underline{\mu}^i_{a,p}(t) \le \tilde{\mu}^i_{a,p}(t) \le \overline{\mu}^i_{a,p}(t)\ \forall t \in \{1, \ldots, N_p(T)\}$$

almost surely.

Let

$$\begin{aligned}
\overline{L}^i_{a,p}(t) &:= \overline{\mu}^i_{a,p}(t) - \tilde{u}_{a,p}(t)\\
\overline{U}^i_{a,p}(t) &:= \overline{\mu}^i_{a,p}(t) + \tilde{u}_{a,p}(t)\\
\underline{L}^i_{a,p}(t) &:= \underline{\mu}^i_{a,p}(t) - \tilde{u}_{a,p}(t)\\
\underline{U}^i_{a,p}(t) &:= \underline{\mu}^i_{a,p}(t) + \tilde{u}_{a,p}(t).
\end{aligned}$$

Note that $\Pr(\mu^i_a(\tilde{x}_p(t)) \notin [L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]) = 0$ for $N_{a,p}(t) = 0$ since we have $L^i_{a,p}(t) = -\infty$ and $U^i_{a,p}(t) = +\infty$ when $N_{a,p}(t) = 0$. Thus, in the rest of the proof, we focus on the case when $N_{a,p}(t) > 0$. It can be shown that

$$\begin{aligned}
&\{\mu^i_a(\tilde{x}_p(t)) \notin [L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]\}\\
&\subset \{\mu^i_a(\tilde{x}_p(t)) \notin [\overline{L}^i_{a,p}(t) - v, \overline{U}^i_{a,p}(t) + v]\}\\
&\cup \{\mu^i_a(\tilde{x}_p(t)) \notin [\underline{L}^i_{a,p}(t) - v, \underline{U}^i_{a,p}(t) + v]\}. \qquad (15)
\end{aligned}$$

The following inequalities can be obtained from the Hölder continuity assumption:

$$\mu^i_a(\tilde{x}_p(t)) \le \overline{\mu}^i_{a,p} \le \mu^i_a(\tilde{x}_p(t)) + L\left(\frac{\sqrt{d}}{m}\right)^\alpha \qquad (16)$$

$$\mu^i_a(\tilde{x}_p(t)) - L\left(\frac{\sqrt{d}}{m}\right)^\alpha \le \underline{\mu}^i_{a,p} \le \mu^i_a(\tilde{x}_p(t)). \qquad (17)$$

Since $v = L(\sqrt{d}/m)^\alpha$, using (16) and (17) it can be shown that

$$\begin{aligned}
(i)\ &\{\mu^i_a(\tilde{x}_p(t)) \notin [\overline{L}^i_{a,p}(t) - v, \overline{U}^i_{a,p}(t) + v]\}\\
&\quad \subset \{\overline{\mu}^i_{a,p} \notin [\overline{L}^i_{a,p}(t), \overline{U}^i_{a,p}(t)]\},\\
(ii)\ &\{\mu^i_a(\tilde{x}_p(t)) \notin [\underline{L}^i_{a,p}(t) - v, \underline{U}^i_{a,p}(t) + v]\}\\
&\quad \subset \{\underline{\mu}^i_{a,p} \notin [\underline{L}^i_{a,p}(t), \underline{U}^i_{a,p}(t)]\}.
\end{aligned}$$

Plugging these into (15), we get

$$\begin{aligned}
&\{\mu^i_a(\tilde{x}_p(t)) \notin [L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]\}\\
&\subset \{\overline{\mu}^i_{a,p} \notin [\overline{L}^i_{a,p}(t), \overline{U}^i_{a,p}(t)]\} \cup \{\underline{\mu}^i_{a,p} \notin [\underline{L}^i_{a,p}(t), \underline{U}^i_{a,p}(t)]\}.
\end{aligned}$$

Then, using the equation above and the union bound, we obtain

$$\begin{aligned}
\Pr(\mathrm{UC}^i_{a,p}) \le \Pr&\left(\bigcup_{t=1}^{N_p(T)} \{\overline{\mu}^i_{a,p} \notin [\overline{L}^i_{a,p}(t), \overline{U}^i_{a,p}(t)]\}\right)\\
&+ \Pr\left(\bigcup_{t=1}^{N_p(T)} \{\underline{\mu}^i_{a,p} \notin [\underline{L}^i_{a,p}(t), \underline{U}^i_{a,p}(t)]\}\right).
\end{aligned}$$

Both terms on the right-hand side of the inequality above can be bounded using the concentration inequality in Appendix D. Using $\delta = 1/(4|\mathcal{A}|m^d T)$ in Appendix D gives

$$\Pr(\mathrm{UC}^i_{a,p}) \le \frac{1}{2|\mathcal{A}|m^d T}$$

since $1 + N_{a,p}(T) \le T$. Then, using the union bound, we obtain

$$\Pr(\mathrm{UC}^i_p) \le \frac{1}{2m^d T}$$

and

$$\Pr(\mathrm{UC}_p) \le \frac{1}{m^d T}.$$

Appendix B
Proof of Lemma 5

Let $\mathcal{T}_{a,p} := \{1 \le l \le N_p(t) : \tilde{a}_p(l) = a\}$ and $\tilde{\mathcal{T}}_{a,p} := \{l \in \mathcal{T}_{a,p} : \tilde{N}_{a,p}(l) \ge 1\}$. By Lemma 2, we have

$$\begin{aligned}
\mathrm{Reg}^1_p(t) &= \sum_{a\in\mathcal{A}} \sum_{l\in\mathcal{T}_{a,p}} \left(\mu^1_*(\tilde{x}_p(l)) - \mu^1_{\tilde{a}_p(l)}(\tilde{x}_p(l))\right)\\
&\le \sum_{a\in\mathcal{A}} \sum_{l\in\tilde{\mathcal{T}}_{a,p}} \left(U^1_{\tilde{a}_p(l),p}(l) - L^1_{\tilde{a}_p(l),p}(l) + 2(\beta+2)v\right)\\
&\quad + |\mathcal{A}|C^1_{\max}\\
&\le \sum_{a\in\mathcal{A}} \sum_{l\in\tilde{\mathcal{T}}_{a,p}} \left(U^1_{\tilde{a}_p(l),p}(l) - L^1_{\tilde{a}_p(l),p}(l)\right)\\
&\quad + 2(\beta+2)vN_p(t) + |\mathcal{A}|C^1_{\max}. \qquad (18)
\end{aligned}$$

We also have

$$
\begin{aligned}
&\sum_{a\in\mathcal{A}}\sum_{l\in\tilde{\mathcal{T}}_{a,p}}\left(U^1_{\tilde{a}_p(l),p}(l)-L^1_{\tilde{a}_p(l),p}(l)\right)\\
&\leq\sum_{a\in\mathcal{A}}\left(B_{m,T}\sum_{l\in\tilde{\mathcal{T}}_{a,p}}\sqrt{\frac{1}{\tilde{N}_{a,p}(l)}}\right)\\
&\leq B_{m,T}\sum_{a\in\mathcal{A}}\sum_{k=0}^{N_{a,p}(t)-1}\sqrt{\frac{1}{1+k}}\\
&\leq 2B_{m,T}\sum_{a\in\mathcal{A}}\sqrt{N_{a,p}(t)} \qquad (19)\\
&\leq 2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)} \qquad (20)
\end{aligned}
$$

where $B_{m,T}=2\sqrt{2A_{m,T}}$, and (19) follows from the fact that

$$
\sum_{k=0}^{N_{a,p}(t)-1}\sqrt{\frac{1}{1+k}}\leq\int_{x=0}^{N_{a,p}(t)}\frac{1}{\sqrt{x}}dx=2\sqrt{N_{a,p}(t)}.
$$

Combining (18) and (20), we obtain that on event UC^c

$$
\mathrm{Reg}^1_p(t)\leq|\mathcal{A}|C^1_{\max}+2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)}+2(\beta+2)vN_p(t).
$$

Appendix C
Proof of Lemma 6

Using the result of Lemma 4, the contribution to the regret of the non-dominant objective in rounds for which $\tilde{u}_{\hat{a}^*_1(t),p}(t)>\beta v$ is bounded by

$$
C^2_{\max}|\mathcal{A}|\left(\frac{2A_{m,T}}{\beta^2 v^2}+1\right). \qquad (21)
$$

Let $\mathcal{T}^2_{a,p}:=\{l\leq N_p(t):\tilde{a}_p(l){=}a$ and $\tilde{N}_{a,p}(l)\geq 2A_{m,T}/(\beta^2 v^2)\}$. By Lemma 3, we have

$$
\begin{aligned}
&\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{T}^2_{a,p}}\left(\mu^2_*(\tilde{x}_p(l))-\mu^2_{\tilde{a}_p(l)}(\tilde{x}_p(l))\right)\\
&\leq\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{T}^2_{a,p}}\left(U^2_{\tilde{a}_p(l),p}(l)-L^2_{\tilde{a}_p(l),p}(l)+2v\right)\\
&\leq\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{T}^2_{a,p}}\left(U^2_{\tilde{a}_p(l),p}(l)-L^2_{\tilde{a}_p(l),p}(l)\right)+2vN_p(t). \qquad (22)
\end{aligned}
$$

We have on event UC^c

$$
\begin{aligned}
&\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{T}^2_{a,p}}\left(U^2_{\tilde{a}_p(l),p}(l)-L^2_{\tilde{a}_p(l),p}(l)\right)\\
&\leq\sum_{a\in\mathcal{A}}\left(B_{m,T}\sum_{l\in\mathcal{T}^2_{a,p}}\sqrt{\frac{1}{\tilde{N}_{a,p}(l)}}\right)\\
&\leq B_{m,T}\sum_{a\in\mathcal{A}}\sum_{k=0}^{N_{a,p}(t)-1}\sqrt{\frac{1}{1+k}}\\
&\leq 2B_{m,T}\sum_{a\in\mathcal{A}}\sqrt{N_{a,p}(t)}\\
&\leq 2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)}. \qquad (23)
\end{aligned}
$$

where $B_{m,T}=2\sqrt{2A_{m,T}}$. Combining (21), (22) and (23), we obtain

$$
\begin{aligned}
\mathrm{Reg}^2_p(t)\leq{}&C^2_{\max}|\mathcal{A}|\left(\frac{2A_{m,T}}{\beta^2 v^2}+1\right)+2vN_p(t)\\
&+2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)}.
\end{aligned}
$$

Appendix D
Concentration Inequality [31], [42]

Consider an arm a for which the rewards of objective i are generated by a process $\{R^i_a(t)\}^T_{t=1}$ with $\mu^i_a=\mathrm{E}[R^i_a(t)]$, 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 $\hat{\mu}_a(T)=\sum^{T-1}_{t=1}\mathrm{I}(a(t)=a)R^i_a(t)/N_a(T)$ for $N_a(T)>0$ and $\hat{\mu}_a(T)=0$ for $N_a(T)=0$. Then, for any $0<\delta<1$ with probability at least $1-\delta$ we have

$$
\begin{aligned}
&|\hat{\mu}_a(T)-\mu_a|\\
&\leq\sqrt{\frac{2}{N_a(T)}\left(1+2\log\left(\frac{(1+N_a(T))^{1/2}}{\delta}\right)\right)}\;\forall T\in\mathbb{N}.
\end{aligned}
$$

References

[1] C. Tekin and E. Turgay, “Multi-objective contextual bandits with a dominant objective,” in *Proc. 27th IEEE Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[2] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 661–670.

[3] J. Xu, T. Xing, and M. van der Schaar, “Personalized course sequence recommendations,” *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5340–5352, Oct. 2016.

[4] L. Song, W. Hsu, J. Xu, and M. 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, May 2016.

[5] Y. Gai, B. Krishnamachari, and R. Jain, “Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation,” in *Proc. IEEE Symp. New Frontiers Dyn. Spectrum*, 2010, pp. 1–9.

[6] Y. Gai and B. Krishnamachari, “Distributed stochastic online learning policies for opportunistic spectrum access,” *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6184–6193, Dec. 2014.

[7] T. Le, C. Szepesvari, and R. Zheng, “Sequential learning for multi-channel wireless network monitoring with channel switching costs,” *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5919–5929, Nov. 2014.

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

[9] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 485–492.

[10] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2533–2568, 2014.

[11] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, "Contextual bandits with linear payoff functions," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 208–214.

[12] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and non-stochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[13] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, vol. 20, pp. 1096–1103.

[14] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1638–1646.

[15] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[16] R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Adv. Appl. Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[17] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2–3, pp. 235–256, 2002.

[18] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 359–376.

[19] C. Tekin, S. Zhang, and M. van der Schaar, "Distributed online learning in social recommender systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 638–652, Aug. 2014.

[20] C. Tekin, J. Yoon, and M. van der Schaar, "Adaptive ensemble learning with confidence bounds," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 888–903, Feb. 2017.

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

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

[23] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini, "Finite-time analysis of kernelised contextual bandits," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, 2013, pp. 654–663.

[24] M. Dudik *et al.*, "Efficient optimal learning for contextual bandits," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, 2011, pp. 169–178.

[25] S. Q. Yahyaa, M. M. Drugan, and B. Manderick, "Knowledge gradient for multi-objective multi-armed bandit algorithms," in *Proc. 6th Int. Conf. Agents Artif. Intell.*, 2014, vol. 1, pp. 74–83.

[26] S. Q. Yahyaa and B. Manderick, "Thompson sampling for multi-objective multi-armed bandits problem," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2015, pp. 47–52.

[27] S. Q Yahyaa, M. M. Drugan, and B. Manderick, "Annealing-Pareto multi-objective multi-armed bandit algorithm," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn.*, 2014, pp. 1–8.

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

[29] Z. Gábor, Z. Kalmár, and C. Szepesvári, "Multi-criteria reinforcement learning," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, vol. 98, pp. 197–205.

[30] S. Mannor and N. Shimkin, "A geometric approach to multi-criterion reinforcement learning," *J. Mach. Learn. Res.*, vol. 5, pp. 325–360, 2004.

[31] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2312–2320.

[32] M. Ehrgott, *Multicriteria Optimization*, vol. 491. New York, NY, USA: Springer, 2005.

[33] S. Sarkar and L. Tassiulas, "Fair allocation of utilities in multirate multicast networks: A framework for unifying diverse fairness objectives," *IEEE Trans. Autom. Control*, vol. 47, no. 6, pp. 931–944, Jun. 2002.

[34] D. T. Hoang, E. L. Linzer, and J. S. Vitter, "Lexicographic bit allocation for MPEG video," *J. Vis. Commun. Image Represent.*, vol. 8, no. 4, pp. 384–404, 1997.

[35] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Yi-Cheng Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proc. National Acad. Sci.*, vol. 107, no. 10, pp. 4511–4515, 2010.

[36] J. A. Konstan, S. M. McNee, C.-N. Ziegler, R. Torres, N. Kapoor, and J. Riedl, "Lessons on applying automated recommender systems to information-seeking tasks," in *Proc. AAAI Conf. Artif. Intell.*, 2006, vol. 6, pp. 1630–1633.

[37] S. McCoy, A. Everard, P. Polak, and D. F. Galletta, "The effects of online advertising," *Commun. ACM*, vol. 50, no. 3, pp. 84–88, 2007.

[38] V. Shah-Mansouri, A.-H. Mohsenian-Rad, and V. WS Wong, "Lexicographically optimal routing for wireless sensor networks with multiple sinks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1490–1500, Mar. 2009.

[39] R. Li, A. Eryilmaz, L. Ying, and N. B. Shroff, "A unified approach to optimizing performance in networks serving heterogeneous flows," *IEEE/ACM Trans. Netw.*, vol. 19, no. 1, pp. 223–236, Feb. 2011.

[40] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and Manfred K Warmuth, "How to use expert advice," *J. ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[41] W. Chu *et al.*, "A case study of behavior-driven conjoint analysis on Yahoo!: Front page today module," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1097–1104.

[42] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Math. Oper. Res.*, vol. 39, no. 4, pp. 1221–1243, 2014.

Cem Tekin (M'13) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011 and 2013, respectively. From February 2013 to January 2015, he was a Postdoctoral Scholar with the University of California, Los Angeles, CA, USA. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey. His research interests include reinforcement learning, multi-armed bandit problems, data mining, multiagent systems, and smart healthcare. He received the University of Michigan Electrical Engineering Departmental Fellowship in 2008, and the Fred W. Ellersick award for the best paper in MILCOM 2009.

Eralp Turğay received the B.Sc. degree in electrical and electronics engineering, in 2015, from Bilkent University, Ankara, Turkey, where he is currently working toward the Master's degree with the Department of Electrical and Electronics Engineering. His research interests include machine learning and multi-armed bandit problems.