ALGORITHMS AND REGRET BOUNDS FOR MULTI-OBJECTIVE CONTEXTUAL BANDITS WITH SIMILARITY INFORMATION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Eralp Turğay

January 2019

Algorithms and Regret Bounds for Multi-objective Contextual Bandits with Similarity Information
By Eralp Turğay
January 2019

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Cem Tekin(Advisor)

Orhan Arıkan

Umut Orguner

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

ALGORITHMS AND REGRET BOUNDS FOR MULTI-OBJECTIVE CONTEXTUAL BANDITS WITH SIMILARITY INFORMATION

Eralp Turğay
M.S. in Electrical and Electronics Engineering
Advisor: Cem Tekin
January 2019

Contextual bandit algorithms have been shown to be effective in solving sequential decision making problems under uncertain environments, ranging from cognitive radio networks to recommender systems to medical diagnosis. Many of these real world applications involve multiple and possibly conflicting objectives. In this thesis, we consider an extension of contextual bandits called multi-objective contextual bandits with similarity information. Unlike single-objective contextual bandits, in which the learner obtains a random scalar reward for each arm it selects, in the multi-objective contextual bandits, 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. For this setting, first, we propose a new multi-objective contextual multi-armed bandit problem with similarity information that has two objectives, where one of the objectives dominates the other objective. Here, 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. 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 evaluate the performance of MOC-MAB in synthetic and real-world datasets. In the next problem, we consider a multi-objective contextual bandit problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set, which is endowed with the similarity information. We propose an online learning algorithm called Pareto Contextual Zooming (PCZ), and prove that it achieves sublinear in the number of rounds Pareto regret, which is near-optimal.

Keywords: Online learning, contextual bandits, multi-objective bandits, dominant objective, multi-dimensional regret, Pareto regret, 2D regret, Similarity Information.

ÖZET

BENZERLİK BİLGİSİNE SAHİP ÇOK AMAÇLI BAĞLAMSAL HAYDUT PROBLEMLERİNDE PİŞMANLIK SINIRLARI VE ALGORİTMALAR

Eralp Turğay
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Cem Tekin
Ocak 2019

Bağlamsal haydut algoritmalarının, bilişsel radyo ağlarından tavsiye sistemlerine ve tıbbi tanıya kadar, belirsiz ortamlarda sıralı karar verme problemlerini çözmede etkili olduğu gösterilmiştir. Bu uygulamaların birçoğu birden fazla ve muhtemelen birbiriyle çelişen amaçlar içerir. Bu tezde, bağlamsal haydut problemlerinin bir uzantısı olan benzerlik bilgisine sahip çok amaçlı bağlamsal haydut problemleri ele alınmıştır. Öğrenicinin seçtiği her kol için rastgele bir skaler ödül aldığı tek amaçlı bağlamsal haydut problemlerinin aksine, çok amaçlı bağlamsal haydut problemlerinde, öğrenici seçtiği her kol için rastgele bir ödül vektörü elde eder. Bu ödül vektörünün her bir elemanı bir amaca karşılık gelir ve ödül vektörünün dağılımı, o turun başlangıcında gözlemlenen bağlama bağlıdır. İlk olarak, bu tezde, bu yapıya uyan, amaçlardan birinin diğer amaca baskın olduğu, iki amaçlı, benzerlik bilgisine sahip yeni bir çok amaçlı bağlamsal haydut problemi tanımlanmıştır. Burada, öğrenicinin amacı, baskın olan amaçtaki toplam ödülünü en üst düzeye çıkardığından emin olmak kaydıyla baskın olmayan amaçtaki toplam ödülünü en üst düzeye çıkarmaktır. Bu problem için bir çok amaçlı bağlamsal haydut algoritması (the multi-objective contextual multi-armed bandit algorithm veya kısaca MOC-MAB) önerilmiştir ve iki farklı performans ölçütü tanımlanmıştır: 2-boyutlu (2D) pişmanlık ve Pareto pişmanlık. Ardından, MOC-MAB'ın hem 2D pişmanlığının hem de Pareto pişmanlığının, tur sayısının altdoğrusal bir fonksiyonu olduğu gösterilmiştir. Ayrıca MOC-MAB'ın sentetik ve gerçek dünya veri kümelerindeki performansı değerlendirilmiştir. Bir sonraki problemde, rastgele sayıda amaca ve benzerlik bilgisine sahip, aynı zamanda yüksek boyutlu ve muhtemelen sayılamayan bir kol kümesi bulunduran çok amaçlı bağlamsal haydut problemi ele alınmıştır. Pareto Contextual Zooming (PCZ) adında bir çevrimiçi öğrenme algoritması önerilmiş ve PCZ'nin Pareto pişmanlığının, tur sayısının altdoğrusal bir fonksiyonu olduğu ve bu fonksiyounun optimuma yakın olduğu gösterilmiştir.

Anahtar sözcükler: Çevrimiçi öğrenme, bağlamsal haydut problemleri, çok amaçlı haydut problemleri, baskın amaç, çok boyutlu pişmanlık, Pareto pişmanlık, 2D pişmanlık, Benzerlik Bilgisi.

Acknowledgement

I would first like to thank my advisor Dr. Cem Tekin, for his support, and guidance throughout my graduate studies. His technical and editorial advice was essential to the completion of this dissertation.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Orhan Arıkan, and Dr. Umut Orguner, for their time, and valuable feedbacks.

I am indebted to Kubilay Ekşioğlu, Cem Bulucu, Ümitcan Şahin, Safa Şahin and Anjum Qureshi for enjoyable coffee breaks, valuable conversations and making my stay in Ankara a pleasant and memorable one.

Finally, I would like to thank my family for all their support they gave me in everything that was accomplished.

This work was supported by TUBITAK Grants 116C043 and 116E229.

Contents

List of Figures

List of Tables

List of Publications

This thesis includes content from following publications:

1. E. Turgay, D. Oner, and C. Tekin, ”Multi-objective contextual bandit problem with similarity information” in *Proc. 21st. Int. Conf. on Artificial Intelligence and Statistics*, pp. 1673-1681, 2018.

2. C. Tekin and E. Turgay ”Multi-objective contextual multi-armed bandit with a dominant objective” *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3799-3813, 2018.

Chapter 1

Introduction

In reinforcement learning, a learner interacts with its environment, and modifies its actions based on feedback received in response to its actions. The standard reinforcement learning framework considers the learner operating in discrete time steps (rounds) and the experience the learner has gathered from interaction with the environment in one round may thus be represented by a set of four-tuples: current state of the environment, action, observed feedback (reward) and next state of the environment. The aim of the learner is to maximize its cumulative reward and this learning model naturally appears in many real-world problems. For instance, an autonomous car receives information about the position and the velocity of its surrounding objects, chooses a direction to move and receives a feedback about how many meters have been moved to the desired location. The exact behavior of the environment is unknown to the learner and it is learned by the aforementioned four-tuple vectors obtained from previous interactions. However, various frameworks such as Markov Decision Processes (MDPs) or Multi-Armed Bandits (MABs) are used to model the environment and in general, it is assumed that the learner knows the framework but not its parameters.

One class of reinforcement learning methods is based on the MAB framework which provides a principled way to model sequential decision making in an uncertain environment. In the classical MAB problem, originally proposed by Robbins [3], a gambler is presented with a sequence of trials where in each trial it has to choose a slot machine (it is also called “one-armed bandit”, and referred to as “arm” hereafter) to play from a set of arms, each providing stochastic rewards over time with unknown distribution. The gambler observes a noisy reward based on its selection and the goal of the gambler is to use the knowledge it obtains through these observations to maximize its expected long-term reward. For this, the gambler 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 (trying out each arm to find the best one) and exploitation (playing the arm believed to have highest expected reward). This sequential intelligent decision making mechanism has received much attention because of the simple model it provides of the trade-off between exploration and exploitation, and consequently has been widely adopted in real-world applications. While these applications ranging from cognitive radio networks [4] to recommender systems [5] to medical diagnosis [6] require intelligent decision making mechanisms that learn from the past, majority of them involve side-observations that can guide the decision making process by informing the learner about the current state of the environment, which does not fit into the classical MAB model. These tasks can be formalized by new MAB models, called contextual multi-armed bandits (contextual MABs), that learn how to act optimally based on side-observations [5, 7, 8]. On the other hand, the aforementioned real-world applications also involve multiple and possibly conflicting objectives. For instance, these objectives include throughput and reliability in a cognitive radio network, semantic match and job-seeking intent in a talent recommender system [9], and sensitivity and specificity in medical diagnosis. This motivates us to work on multi-objective contextual MAB problems which address the learning challenges that arise from side-observations and presence of multiple objectives at the same time.

In the multi-objective contextual MAB problems, at the beginning of each round, the learner observes a context from a context set and then selects an arm from an arm set. In general, size of the context set is infinite but size of the arm set can be finite or infinite, depending on the application area of the contextual bandit model. At the end of the round, the learner receives a multi-dimensional reward whose distribution depends on the observed context and the selected arm. Aim of the learner is to maximize its total reward for each objective. However, since the rewards are no longer scalar, the definition of a benchmark to compare the learner against becomes obscure. Different performance metrics are proposed such as Pareto regret and scalarized regret [10]. Pareto regret measures sum of the distances of the arms selected by the learner to the Pareto front. On the other hand, in the scalarized approach, weights are assigned for each objective, from which for each arm a weighted sum of the expected rewards of the objectives are calculated and the difference between the optimal arm and the selected arm is defined as the scalarized regret. However, these performance metrics cannot model all existing real-world problems. For instance, consider a multichannel communication system, where a user chooses a channel and a transmission rate at each round and when the user completes its transmission at the end of a round, it receives a two dimensional reward vector that contains throughput and reliability of the transmission. Aim of the user is to choose the channel that maximizes reliability among all channels that maximize throughput. To model this problem accurately, dominance relation between the objectives should be considered.

In this thesis, we consider two multi-objective contextual MAB problems. In the first one, we work on multi-objective contextual MAB problem with similarity information that has two objectives, where one of the objectives dominates the other objective. Simply, similarity information is an assumption that relates the distances between contexts to the distances between expected rewards of an arm. For this problem, we use a novel performance metric, called the 2D regret, which we proposed in [11] to deal with problems involving dominant and non-dominant objectives. In the second problem, we consider a multi-objective contextual MAB problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set, which is also endowed with the similarity information. Additionally, we include the proposed solutions for these problems in [11] and [12].

Essentially, the first problem is a multi-objective contextual MAB with two objectives. We assume that the learner seeks to maximize its expected reward in the second (non-dominant) objective subject to the constraint that it maximizes its expected reward in the first (dominant) objective. We call this problem multi-objective contextual multi-armed bandit with a dominant objective (CMAB-DO). In this problem, we assume that the learner is endowed with similarity information, which relates the variation in the expected reward of an arm as a function of the context to the distance between the contexts. It is a common assumption in contextual MAB literature [8,13,14], and merely states that the expected reward function is Lipschitz continuous in the context. In CMAB-DO, the learner competes with the optimal arm (i.e., the arm that maximizes the expected reward in the second objective among all arms that maximize the expected reward in the first objective), and hence, the performance of the learner is measured in terms of its 2D regret which 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 round T. For this problem, we propose an online learning algorithm called multi-objective contextual multi armed bandit algorithm (MOC-MAB) and we prove that it achieves $\tilde{O}(T^{(2\alpha+d_x)/(3\alpha+d_x)})$ 2D regret, where d_x is the dimension of the context and α is a constant that depends on the similarity information. Hence, MOC-MAB is average-reward optimal in the limit $T \to \infty$ in both objectives. Additionally, we show that MOC-MAB achieves $\tilde{O}(T^{(2\alpha+d)/(3\alpha+d)})$ Pareto regret, since the optimal arm lies in the Pareto front. We also show that by adjusting the parameters, MOC-MAB can achieve $\tilde{O}(T^{(\alpha+d_x)/(2\alpha+d_x)})$ Pareto regret such that it becomes order optimal up to a logarithmic factor [8] but this comes at an expense of making the regret in the non-dominant objective of MOC-MAB linear in the number of rounds. Performance of MOC-MAB is evaluated through simulations and it is observed that the proposed algorithm outperforms its competitors, which are not specifically designed to deal with problems involving dominant and non-dominant objectives.

In the next problem, we consider a multi-objective contextual MAB problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set (also called multi-objective contextual $\mathcal{X}$-armed bandit problem). In this problem, since the arm set may contain infinite number of arms, it is impossible to explore all arms and find the optimal one. To facilitate the learning in the arm and the context sets, we assume that the learner is endowed with similarity information in these sets such that it relates the distances between context-arm pairs to the distances between expected rewards of these pairs. This similarity information is an intrinsic property of the similarity space, which consists of all feasible context-arm pairs, and implies that the expected reward function is Lipschitz continuous.

In order to evaluate the performance of the algorithms in this problem, we adopt the notion of contextual Pareto Regret which we defined in [11] for two objectives, and extend it to work for an arbitrary number of objectives. However, the challenge we faced in this problem is Pareto front can vary from context to context, which makes its complete characterization difficult even when the expected rewards of the context-arm pairs are known. Additionally, in many applications where sacrificing one objective over another one is disadvantageous, it is necessary to ensure that all of the Pareto optimal context-arm pairs are equally treated. We address these challenges and propose an online learning algorithm called Pareto Contextual Zooming (PCZ). We also show that it achieves $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ Pareto regret, where d_p is the Pareto zooming dimension, which is an optimistic version of the covering dimension that depends on the size of the set of near-optimal context-arm pairs. PCZ is built on the contextual zooming algorithm in [7], and achieves this regret bound by adaptively partitioning context-arm set according to context arrivals, empirical distribution of arm selections and observed rewards. We also find a lower bound $\Omega(T^{(1+d_p)/(2+d_p)})$ for this problem in [12], so Pareto regret of PCZ is order optimal up to a logarithmic factor.

1.1 Applications of Multi-objective Contextual Bandits in Cognitive Radio Networks

In this section, we describe potential applications of the multi-objective contextual MAB for Cognitive Radio Networks (CRNs). Simply, a CRN is a wireless communication system that adopts and optimizes its transmission parameters according to the changes in its surroundings. In that way, it improves the utilization efficiency of the existing radio spectrum. CRNs include the methods that optimize the inter-layer and inter-user communication parameters and actions, and many of these methods use a MAB framework [15–17]. However, many of the problems in CRNs also involve multiple and possibly conflicting objectives. Hence, multi-objective contextual MAB can be adopted for these problems, and three examples of such applications are described below.

1.1.1 Multichannel Communication

Consider a multi-channel 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 $x_t := \{\mathrm{SNR}_{Q,t}\}_{Q \in \mathcal{Q}}$, where $\mathrm{SNR}_{Q,t}$ 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 $\mathrm{SNR}_{Q,t}$ 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 one of the objectives is related to throughput and the other one is related to reliability. Here, in the objective related to throughput, the learner receives "0" reward for failed transmission and "1" reward for successful transmission. In the other objective, if the transmission is successful, the learner receives a reward directly proportional to the selected transmission rate. It is usually the case that the probability of failed transmission increases with the transmission rate, so maximizing the throughput and reliability are usually conflicting objectives. This problem is adopted for the multi-objective contextual MAB with a dominant objective setting and details of the application are given in Section 3.1. Additionally, illustrative results on this application are given in Section 3.5.

1.1.2 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. 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 multi-objective contextual MAB. 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 and delay over the selected path. These costs are converted to rewards by extracting them from a constant value. This problem is also adopted for the multi-objective contextual MAB with a dominant objective setting and details of the application are given in Section 3.1.

1.1.3 Cross-layer Learning in Heterogeneous Cognitive Radio Networks

In [18], a contextual MAB model for cross-layer learning in heterogeneous cognitive radio networks is proposed. In this method, in the physical layer, application adaptive modulation (AAM) is implemented and bit error rate constraint is considered as the context for the contextual MAB model. It is assumed that the channel state information is not known beforehand. Bit error rate constraint is given to the physical layer from the application layer and since each application has a dynamic packet error rate constraint, it is used to determine the bit error rate constraint at the physical layer. However, this problem intrinsically contains two different objectives and it can be well modeled by the multi-objective contextual MAB framework. One of the objectives is to satisfy the bit error rate constraint and the other one is to maximize the expected bits per symbol (BPS) rate. At the beginning of each round, the learner observes a context that informs the learner about the bit error rate constraint, then it selects an AAM from the available set of AAM (For instance, AAM set may correspond to a set of uncoded QAM modulations with different constellation sizes). After this selection, the learner observes a two dimensional reward vector. One of the dimensions of the reward vector corresponds to the bit error rate constraint. For this dimension, the learner receives “1” reward if this constraint is satisfied and, the learner observes “0” reward, if it is not satisfied. The other dimension of the reward vector is directly proportional to the selected BPS rate of the transmission.

In this section, we described potential applications of the multi-objective contextual MAB for CRNs. However, it is also applicable in many other areas such as online binary classification problems and recommender systems. Example applications for each of these problems are given in Section 3.1.

1.2 Our Contributions

Contributions of this thesis are summarized as follows:

- In multi-objective contextual multi-armed bandit with a dominant objective;
    - To the best of our knowledge, our work [11] (which is the extended version of [19]) is the first to consider a multi-objective contextual MAB problem where the expected arm rewards and contexts are related through similarity information.

  - We propose a novel contextual MAB problem with two objectives in which one objective is dominant and the other is non-dominant.
  - To measure the performance of the algorithms in this problem, we propose a new performance metric called 2D regret.
  - We extend the Pareto regret proposed in [10] to take into account the dependence of the Pareto front on the context.
  - We propose a multi-objective contextual multi-armed bandit algorithm (MOC-MAB).
  - We show that both the 2D regret and the Pareto regret of MOC-MAB are sublinear in the number of rounds.
  - We investigate the performance of MOC-MAB numerically.

- In multi-objective contextual $\mathcal{X}$-armed bandit problem;
  - We adopted the notion of contextual Pareto regret defined in [11] for two objectives, and extend it to work for an arbitrary number of objectives.
  - We propose an online learning algorithm called Pareto Contextual Zooming (PCZ).
  - We show that the Pareto regret of PCZ is sublinear in the number of rounds.
  - We show an almost matching lower bound, which shows that our bound is tight up to logarithmic factors.
  - We investigate the performance of PCZ numerically.

1.3 Organization of the Thesis

The rest of the thesis is organized as follows. Next chapter includes literature review. In the first section of Chapter 2, we introduce classical multi-armed problem and then, in Sections 2.2 and 2.3, we include literature review on contextual multi-armed bandits and multi-objective bandits respectively.

In Chapter 3, details of CMAB-DO and MOC-MAB are given. Problem formulation of CMAB-DO, 2D regret and the Pareto regret are described in Section 3.1. We introduce MOC-MAB in Section 3.2 and analyze its regret in Section 3.3. An extension of MOC-MAB that deals with dynamically changing reward distributions is proposed and the case where there are more than two objectives is considered in Section 3.4. Numerical results related to MOC-MAB are presented in Section 3.5. In Chapter 4, we introduce a multi-objective contextual MAB problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set and the learning algorithm that solves this problem, i.e., PCZ. Problem formulation is given in Section 4.1, and PCZ is explained in Section 4.2. Pareto Regret of PCZ is upper bounded in Section 4.3. A lower bound on the Pareto regret of PCZ is given in Section 4.4. Numerical experiments for PCZ are given in Section 4.5. The last chapter concludes the thesis.

Chapter 2

Related Work

In the past decade, many variants of the classical MAB have been introduced. Two notable examples are contextual MAB [7, 21, 22] and multi-objective MAB [10]. Mostly, these examples have been studied separately in prior works, but in our works [11, 12], we fused contextual MAB and multi-objective MAB together due to its applicability in various fields mentioned in Section 1.1 and Section 3.1. Below, we discuss the related work on the classical MAB, contextual MAB, and multi-objective MAB. The differences between our works and related works are summarized in Table 2.1.

2.1 The Classical MAB

This Section was published in [11].[1]

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

[1] ©2018 IEEE. Reprinted, with permission, from C. Tekin and E. Turğay, "Multi-objective Contextual Multi-armed Bandit With a Dominant Objective", *IEEE Transactions on Signal Processing*, July 2018.

Table 2.1: Comparison of the regret bounds and assumptions in our work with the related literature.

Bandit algorithm	Regret bound	Multi-objective	Contextual	Linear rewards	Similarity assumption
Contextual Zooming [7]	$\tilde{O}(T^{1-1/(2+d_z)})$	No	Yes	No	Yes
Query-Ad-Clustering [8]	$\tilde{O}(T^{1-1/(2+d_c)})$	No	Yes	No	Yes
SupLinUCB [20]	$\tilde{O}(\sqrt{T})$	No	Yes	Yes	No
Pareto-UCB1 [10]	$O(\log(T))$	Yes	No	No	No
Scalarized-UCB1 [10]	$O(\log(T))$	Yes	No	No	No
PCZ [12](our work)	$\tilde{O}(T^{1-1/(2+d_p)})$ (Pareto regret)	Yes	Yes	No	Yes
MOC-MAB [11] (our work)	$\tilde{O}(T^{(2\alpha+d_x)/(3\alpha+d_x)})$ (2D and Pareto regrets) $\tilde{O}(T^{(\alpha+d_x)/(2\alpha+d_x)})$ (Pareto regret only)	Yes	Yes	No	Yes

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 [23], 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 [24] 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 [25]. 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 [26].

2.2 The Contextual MAB

This Section was published in [11].[2]

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 [27], medical diagnosis [14] and big data stream mining [13]. 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 [8], 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 space into disjoint sets and to estimate the expected arm rewards for each set in the partition separately. A parallel work [7] 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

[2] ©2018 IEEE. Reprinted, with permission, from C. Tekin and E. Turğay, "Multi-objective Contextual Multi-armed Bandit With a Dominant Objective", *IEEE Transactions on Signal Processing*, July 2018

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 [28], 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 [5]. A modified version of this algorithm, named SupLinUCB, is studied in [20], and is shown to achieve $\tilde{O}(\sqrt{Td})$ regret, where d is the dimension of the context. Another work [29] considers LinUCB and SupLinUCB with kernel functions and proposes an algorithm whwith $\tilde{O}(\sqrt{T\tilde{d}})$ 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 [21], and more efficient learning algorithms with $\tilde{O}(T^{1/2})$ regret are developed in [30] and [22].

Our problems in this thesis are similar to the problems in the first category in terms of the context arrivals and existence of the similarity information.

2.3 The Multi-objective MAB

This Section was published in [11].[3]

In the multi-objective MAB problem, the learner receives a multi-dimensional

[3] ©2018 IEEE. Reprinted, with permission, from C. Tekin and E. Turğay, "Multi-objective Contextual Multi-armed Bandit With a Dominant Objective", *IEEE Transactions on Signal Processing*, July 2018

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: 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 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 [10], 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 [31]. Numerous other variants are also considered in prior works, including the Pareto Thompson sampling algorithm in [32] and the Annealing Pareto algorithm in [33].

On the other hand, in the scalarized approach [10, 34], 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 [10] achieves $O(S' \log(T/S'))$ scalarized regret where S' is the number of scalarization functions used by the algorithm.

In addition to the works mentioned above, several other works consider multi-criteria reinforcement learning problems, where the rewards are vector-valued [35, 36].

Chapter 3

Multi-objective Contextual Multi-Armed Bandit with a Dominant Objective

In this chapter, we consider a multi-objective contextual MAB with two objectives, where one of the objectives dominates the other objective. We call this problem contextual multi-armed bandit with a dominant objective (CMAB-DO). For this problem, we define two performance measures: the 2-dimensional (2D) regret and the Pareto regret. The first section includes problem formulation of CMAB-DO, 2D regret and the Pareto regret definitions. Then, we propose MOC-MAB in Section 3.2 and show that both the 2D regret and the Pareto regret of MOC-MAB are sublinear in the number of rounds in Section 3.3. An extension of MOC-MAB that deals with dynamically changing reward distributions is proposed and the case where there are more than two objectives is considered in Section 3.4 and we present numerical results of MOC-MAB in Section 3.5. This work was published in [11] .[1]

[1] ©2018 IEEE. Reprinted, with permission, from C. Tekin and E. Turğay, ”Multi-objective Contextual Multi-armed Bandit With a Dominant Objective”, *IEEE Transactions on Signal Processing*, July 2018

3.1 Problem Formulation

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_x-dimensional context denoted by x_t. Without loss of generality, we assume that x_t lies in the context set $\mathcal{X} := [0, 1]^{d_x}$. 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,[2] i.e.,

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

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 that is defined above coincides with lexicographic optimality [37], which is widely used in multicriteria optimization, and has been considered in numerous applications such as achieving fairness in multirate multicast networks [38] and bit allocation for MPEG video coding [39].

We assume that the expected rewards are Hölder continuous in the context, which is a common assumption in the contextual bandit literature [8, 13, 14].

[2] 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 [1]. 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.

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*

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

Since Hölder continuity implies continuity, for any nontrivial 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 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 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 $a^*(x)$.*

3.1.1 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

$$\text{Reg}^i(T) := \sum_{t=1}^{T} \mu_*^i(x_t) - \sum_{t=1}^{T} \mu_{a_t}^i(x_t), \ i \in \{1, 2\} \tag{3.1}$$

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 [10], 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,*

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

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

$$\text{PR}(T) := \sum_{t=1}^{T} \Delta_{a_t}(x_t). \tag{3.2}$$

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.

3.1.2 Applications of CMAB-DO

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

3.1.2.1 Multichannel Communication

Consider a multi-channel 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 $x_t := \{\mathrm{SNR}_{Q,t}\}_{Q \in \mathcal{Q}}$, where $\mathrm{SNR}_{Q,t}$ 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 $\mathrm{SNR}_{Q,t}$ 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 $\mu^2_{a_{R,Q}}(x_t) = 1 - p_{\mathrm{out}}(R, Q, x_t)$, where $p_{\mathrm{out}}(\cdot)$ denotes the outage probability. Here, $p_{\mathrm{out}}(R, Q, x_t)$ 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 $\mu^1_{a_{R,Q}}(x_t) = R(1 - p_{\mathrm{out}}(R, Q, x_t))/R_{\max}$, 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.[3] Illustrative results on this application are given in Section 3.5.

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

3.1.2.2 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.

3.1.2.3 Recommender System

Recommender systems involve optimization of multiple metrics like novelty and diversity in addition to accuracy [40,41]. 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.[4] 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.[5] Thus, $r_t^1 = 1$ if the user clicks to the item and 0 otherwise. Similarly, $r_t^2 = 1$ if the user rates the item as novel and 0 otherwise. The distribution of (r_t^1, r_t^2) depends on x_t and is unknown to the recommender system.

Another closely related application is display advertising [42], where an advertiser can place an ad to the publisher's website for the user currently visiting

[4] An example recommender system that uses this kind of feedback is given in [41].

[5] 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.

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 3.5.

3.1.2.4 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 [43] to obtain routing flows in a wireless sensor network with energy limited nodes. Moreover, [44] 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.

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_t^1 and delay c_t^2 over the selected path. These costs are converted to rewards by setting $r_t^1 = 1 - c_t^1$ and $r_t^2 = 1 - c_t^2$.

3.2 Multi-objective Contextual Multi-armed Bandit Algorithm (MOC-MAB)

We introduce MOC-MAB in this section. Its pseudocode is given in Algorithm 1.

MOC-MAB uniformly partitions $\mathcal{X}$ into m^{d_x} 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.

At round t, MOC-MAB first identifies the hypercube in $\mathcal{P}$ that contains x_t, which is denoted by p^*.[6] Then, it calculates the following indices for the rewards in the dominant and the non-dominant objectives:

$$g^i_{a,p^*} := \hat{\mu}^i_{a,p^*} + u_{a,p^*}, \ i \in \{1, 2\} \tag{3.3}$$

where the *uncertainty level* $u_{a,p} := \sqrt{2A_{m,T}/N_{a,p}}$, $A_{m,T} := (1 + 2\log(4|\mathcal{A}|m^{d_x}T^{3/2}))$ represents the uncertainty over the sample mean estimate of the reward due to the number of instances that are used to compute $\hat{\mu}^i_{a,p^*}$.[7] Hence, a UCB for $\mu^i_a(x)$ is $g^i_{a,p} + v$ for $x \in p$, where $v := Ld_x^{\alpha/2}m^{-\alpha}$ denotes the

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

[7] 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 [45] and [14] for a discussion on the doubling-trick.

Algorithm 1 MOC-MAB

```
Input: $T$, $d_x$, $L$, $\alpha$, $m$, $\beta$
Initialize sets: Create partition $\mathcal{P}$ of $\mathcal{X}$ into $m^{d_x}$ 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 (3.3)
    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 $\hat{\mathcal{A}}^*$ given in (3.4)
        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
```

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 $u_{a^*_1,p^*} \leq \beta v$, where $\beta > 0$ is a constant. For this, it calculates the set of candidate optimal arms given as

$$\begin{aligned}\hat{\mathcal{A}}^* &:= \left\{ a \in \mathcal{A} : g^1_{a,p*} \geq \hat{\mu}^1_{a^*_1,p^*} - u_{a^*_1,p^*} - 2v \right\} \\ &= \left\{ a \in \mathcal{A} : \hat{\mu}^1_{a,p*} \geq \hat{\mu}^1_{a^*_1,p^*} - u_{a^*_1,p^*} - u_{a,p^*} - 2v \right\}.\end{aligned} \tag{3.4}$$

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 $O(d_x)$ 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 $O(d_xT) + O(|\mathcal{A}|T)$ computations in T rounds. In addition, the memory complexity of MOC-MAB is $O(m^{d_x}|\mathcal{A}|)$.*

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 3.5.*

3.3 Regret Analysis of MOC-MAB

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 $\mathcal{H}$, let $\mathcal{H}^c$ denote the complement of that event. For all the parameters defined in Section 3.2, 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 the 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 at 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 $\mathcal{T}_p := \{t \in \{1, \ldots, T\} : x_t \in p\}$ 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

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

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

$$\mathrm{Reg}^i(T) = \sum_{p \in \mathcal{P}} \mathrm{Reg}^i_p(T). \tag{3.5}$$

Thus, the expected regret in objective i becomes

$$\mathrm{E}[\mathrm{Reg}^i(T)] = \sum_{p \in \mathcal{P}} \mathrm{E}[\mathrm{Reg}^i_p(T)]. \tag{3.6}$$

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:

$$\begin{aligned} \mathrm{E}[\mathrm{Reg}^i_p(T)] &= \mathrm{E}[\mathrm{Reg}^i_p(T) \mid \mathrm{UC}] \Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i_p(T) \mid \mathrm{UC}^c] \Pr(\mathrm{UC}^c) \\ &\leq C^i_{\max} N_p(T) \Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{Reg}^i_p(T) \mid \mathrm{UC}^c] \end{aligned} \tag{3.7}$$

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 (3.7), we proceed by bounding the terms in (3.7). For this, we first bound $\Pr(\mathrm{UC}_p)$ in the next lemma.

Lemma 1. *For any $p \in \mathcal{P}$, we have* $\Pr(UC_p) \leq 1/(m^{d_x} T)$.

Proof. From the definitions of $L^i_{a,p}(t)$, $U^i_{a,p}(t)$ and $\mathrm{UC}^i_{a,p}$, it can be observed that the event $\mathrm{UC}^i_{a,p}$ 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 A, 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 $\tilde{R}^i_{a,p}(t)$, $t \in \{1, \dots, N_p(T)\}$ different.

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

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

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

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

almost surely.

Let

$$\overline{L}^i_{a,p}(t) := \overline{\mu}^i_{a,p}(t) - \tilde{u}_{a,p}(t)$$

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

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

$$\begin{aligned}\{\mu^i_a(\tilde{x}_p(t)) \notin [L^i_{a,p}(t) - v, U^i_{a,p}(t) + v]\} \subset \{\mu^i_a(\tilde{x}_p(t)) \notin [\overline{L}^i_{a,p}(t) - v, \overline{U}^i_{a,p}(t) + v]\}\\ \cup \{\mu^i_a(\tilde{x}_p(t)) \notin [\underline{L}^i_{a,p}(t) - v, \underline{U}^i_{a,p}(t) + v]\}.\end{aligned} \tag{3.8}$$

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

$$\mu^i_a(\tilde{x}_p(t)) \leq \overline{\mu}^i_{a,p} \leq \mu^i_a(\tilde{x}_p(t)) + L\left(\frac{\sqrt{d_x}}{m}\right)^\alpha \tag{3.9}$$

$$\mu^i_a(\tilde{x}_p(t)) - L\left(\frac{\sqrt{d_x}}{m}\right)^\alpha \leq \underline{\mu}^i_{a,p} \leq \mu^i_a(\tilde{x}_p(t)). \tag{3.10}$$

Since $v = L\left(\sqrt{d_x}/m\right)^\alpha$, using (3.9) and (3.10) it can be shown that

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

Plugging these into (3.8), 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

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

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

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

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

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

and

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

□

Using the result of Lemma 1, we obtain

$$\Pr(\mathrm{UC}) \leq 1/T \text{ and } \Pr(\mathrm{UC}^c) \geq 1 - 1/T. \tag{3.11}$$

To prove the lemma above, we use the concentration inequality given in Lemma 6 in [1] 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 [14] 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_p^c 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_p^c, we have*

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

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

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

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

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

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

Thus, for both cases, we have

$$U^1_{\tilde{a}(t),p}(t) \geq U^1_{\hat{a}_1^*(t),p}(t) - 2(\beta+1)v \tag{3.12}$$

and

$$U^1_{\hat{a}_1^*(t),p}(t) \geq U^1_{a^*(t),p}(t). \tag{3.13}$$

On event UC_p^c, we also have

$$\mu^1_{a^*(t)}(\tilde{x}_p(t)) \leq U^1_{a^*(t),p}(t) + v \tag{3.14}$$

and

$$\mu^1_{\tilde{a}(t)}(\tilde{x}_p(t)) \geq L^1_{\tilde{a}(t),p}(t) - v. \tag{3.15}$$

By combining (3.12)-(3.15), we obtain

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

□

The lemma below bounds the regret incurred in a round $\tau_p(t)$ for the non-dominant objective on event UC_p^c 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_p^c, for $t \in \{1, \ldots, N_p(T)\}$ if*

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

holds, then we have

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

Proof. When $\tilde{u}_{\hat{a}_1^*(t),p}(t) \leq \beta v$ 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_p^c, 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_p^c holds, we get

$$\mu^2_{\tilde{a}(t)}(\tilde{x}_p(t)) \geq L^2_{\tilde{a}(t),p}(t) - v \tag{3.16}$$

and

$$\mu^2_{a^*(t)}(\tilde{x}_p(t)) \leq U^2_{a^*(t),p}(t) + v \leq U^2_{\tilde{a}(t),p}(t) + v. \tag{3.17}$$

Finally, by combining (3.16) and (3.17), 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 $\tilde{u}_{\hat{a}^*_1(t),p}(t) > \beta v$, $t \in \{1, \dots, N_p(T)\}$.

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 $p \in \mathcal{P}$*

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

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

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

$$\begin{aligned}
\mathrm{Reg}^1_p(t) &= \sum_{a \in \mathcal{A}} \sum_{l \in \mathcal{T}_{a,p}} \left(\mu^1_*(\tilde{x}_p(l)) - \mu^1_{\tilde{a}_p(l)}(\tilde{x}_p(l)) \right) \\
&\leq \sum_{a \in \mathcal{A}} \sum_{l \in \tilde{\mathcal{T}}_{a,p}} \left(U^1_{\tilde{a}_p(l),p}(l) - L^1_{\tilde{a}_p(l),p}(l) + 2(\beta + 2)v \right) + |\mathcal{A}| C^1_{\max} \\
&\leq \sum_{a \in \mathcal{A}} \sum_{l \in \tilde{\mathcal{T}}_{a,p}} \left(U^1_{\tilde{a}_p(l),p}(l) - L^1_{\tilde{a}_p(l),p}(l) \right) + 2(\beta + 2) v N_p(t) + |\mathcal{A}| C^1_{\max}. \quad (3.18)
\end{aligned}$$

We also have

$$\begin{aligned}\sum_{a\in\mathcal{A}}\sum_{l\in\tilde{\mathcal{T}}_{a,p}}\left(U^1_{\tilde{a}_p(l),p}(l)-L^1_{\tilde{a}_p(l),p}(l)\right) &\leq \sum_{a\in\mathcal{A}}\left(B_{m,T}\sum_{l\in\tilde{\mathcal{T}}_{a,p}}\sqrt{\frac{1}{\tilde{N}_{a,p}(l)}}\right)\\ &\leq B_{m,T}\sum_{a\in\mathcal{A}}\sum_{k=0}^{N_{a,p}(t)-1}\sqrt{\frac{1}{1+k}}\\ &\leq 2B_{m,T}\sum_{a\in\mathcal{A}}\sqrt{N_{a,p}(t)} \qquad (3.19)\\ &\leq 2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)} \qquad (3.20)\end{aligned}$$

where $B_{m,T}=2\sqrt{2A_{m,T}}$, and (3.19) 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 (3.18) and (3.20), we obtain that on event UC^c

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

□

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

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

Proof. 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

$$C^2_{\max}|\mathcal{A}|\left(\frac{2A_{m,T}}{\beta^2v^2}+1\right). \qquad (3.21)$$

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

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

(3.22)

We have on event UC^c

$$\begin{aligned}\sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{T}^2_{a,p}}\left(U^2_{\tilde{a}_p(l),p}(l) - L^2_{\tilde{a}_p(l),p}(l)\right) &\le \sum_{a\in\mathcal{A}}\left(B_{m,T}\sum_{l\in\mathcal{T}^2_{a,p}}\sqrt{\frac{1}{\tilde{N}_{a,p}(l)}}\right)\\ &\le B_{m,T}\sum_{a\in\mathcal{A}}\sum_{k=0}^{N_{a,p}(t)-1}\sqrt{\frac{1}{1+k}}\\ &\le 2B_{m,T}\sum_{a\in\mathcal{A}}\sqrt{N_{a,p}(t)}\\ &\le 2B_{m,T}\sqrt{|\mathcal{A}|N_p(t)}. \end{aligned} \tag{3.23}$$

where $B_{m,T} = 2\sqrt{2A_{m,T}}$. Combining (3.21), (3.22) and (3.23), we obtain

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

□

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

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

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

where

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

and

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

Proof. By (3.5) and Lemmas 5 and 6, we have on event UC^c:

$$\begin{aligned}\mathrm{Reg}^1(t) \le& m^{d_x}|\mathcal{A}|C^1_{\max} + 2B_{m,T}\sum_{p\in\mathcal{P}}\sqrt{|\mathcal{A}|N_p(t)} + 2(\beta+2)vt\\ \le& m^{d_x}|\mathcal{A}|C^1_{\max} + 2B_{m,T}\sqrt{|\mathcal{A}|m^{d_x}t} + 2(\beta+2)vt.\end{aligned}$$

and

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

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 $\tilde{O}(T^{\frac{2\alpha+d_x}{3\alpha+d_x}})$.

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

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

and

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

Proof. $\mathrm{E}[\mathrm{Reg}^i(T)]$ is bounded by using the result of Theorem 1 and (3.7):

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

Therefore, we have

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

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

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

and

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

□

From the results above we conclude that both regrets are $\tilde{O}(T^{(2\alpha+d_x)/(3\alpha+d_x)})$, 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 (3.2) is also $\tilde{O}(T^{(2\alpha+d_x)/(3\alpha+d_x)})$.

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

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

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

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

Proof. Consider any $p \in \mathcal{P}$ and $t \in \{1, \dots, N_p(T)\}$. By definition $\Delta_{\tilde{a}(t)}(\tilde{x}_p(t)) \leq \mu^1_{a^*(t)}(\tilde{x}_p(t)) - \mu^1_{\tilde{a}(t)}(\tilde{x}_p(t))$. 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

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

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 $\mathrm{PR}_p(T)$. □

Theorem 3 shows 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 $\tilde{O}(T^{(2\alpha+d_x)/(3\alpha+d_x)})$ when it is run with $m = \lceil T^{1/(3\alpha+d_x)} \rceil$. 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 $m = \lceil T^{1/(2\alpha+d_x)} \rceil$. This will make the Pareto regret $\tilde{O}(T^{(\alpha+d_x)/(2\alpha+d_x)})$ (which matches with the lower bound in [8] 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.

3.4 Extensions

3.4.1 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_x-dimensional context $x_t =$

$(x_{1,t}, x_{2,t}, ..., x_{d_x,t})$ received at round t let $\hat{x}_t := (x_{1,t}, x_{2,t}, ..., x_{d_x+1,t})$ denote the extended context where $x_{d_x+1,t} := (t \mod T_s)/T_s$ is the time context. Let $\hat{\mathcal{X}}$ denote the $d_x + 1$ 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 $0 < \hat{\alpha} \leq 1$ such that for all $i \in \{1, 2\}$ and $a \in \mathcal{A}$, we have*

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

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

$$|\mu_a^i(\hat{x}) - \mu_a^i(\hat{x}')| \leq \hat{L}|x_{d_x+1} - x'_{d_x+1}|^{\hat{\alpha}}$$

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 $\hat{L}$, $\hat{\alpha}$, $m = \lceil T^{1/(3\hat{\alpha}+d_x+1)} \rceil$, and $\beta > 0$ by using the extended context set $\hat{\mathcal{X}}$ instead of the original context set $\mathcal{X}$, we have*

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

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

3.4.2 Lexicographic Optimality for $d_r > 2$ Objectives

Our problem formulation can be generalized to handle $d_r > 2$ objectives as follows. Let $r_t := (r_t^1, \ldots, r_t^{d_r})$ 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 $\mu^i_a(x) > \mu^i_{a'}(x)$, where $i := \min\{k \le j : \mu^k_a(x) \neq \mu^k_{a'}(x)\}$.[8] 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:

$$\begin{aligned}&\mathbf{Reg}(T) := (\text{Reg}^1(T), \ldots, \text{Reg}^{d_r}(T)) \text{ where}\\&\text{Reg}^i(T) := \sum_{t=1}^{T} \mu^i_*(x_t) - \sum_{t=1}^{T} \mu^i_{a_t}(x_t), i \in \{1, \ldots, d_r\}.\end{aligned}$$

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 3.4. 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 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 $\mu^1_a(x) = \mu^1_b(x)$ and $\mu^2_a(x) = \mu^2_b(x)$. 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 $\mu^1_a(x) = \mu^1_b(x)$.

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

3.5 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 [10].
Scalarized UCB1 (S-UCB1): This is the Scalarized Multi-objective UCB1 algorithm proposed in [10].
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 [25], 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 all 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.

3.5.1 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. 3.1 show the change in the regret of the algorithms in both objectives as a 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.

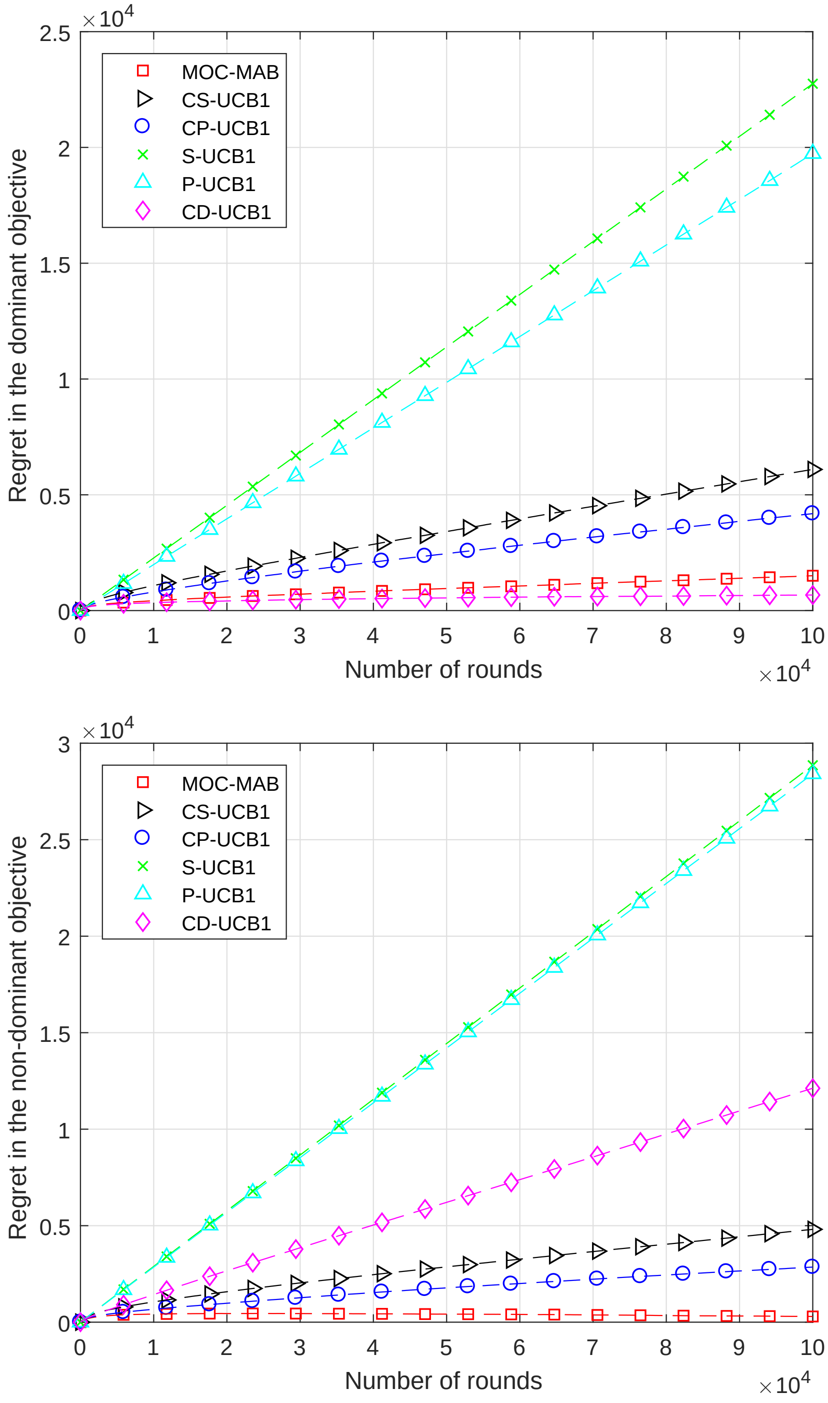

Figure 3.1: Regrets of MOC-MAB and the other algorithms for Experiment 1.

3.5.2 Experiment 2 - Multichannel Communication

In this experiment, we consider the multichannel communication application given in Section 3.1 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. $\mathrm{SNR}_{Q,t}$ 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 $\log_2(1 + h^2_{Q,t}\mathrm{SNR}_{Q,t}) < R$.

Every algorithm is run 20 times and the results are averaged over these runs. Simulation results are given in Fig.3.2 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.

3.5.3 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 $x_t = (x^{\mathrm{usr}}_t, x^{\mathrm{ad}}_t)$. 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

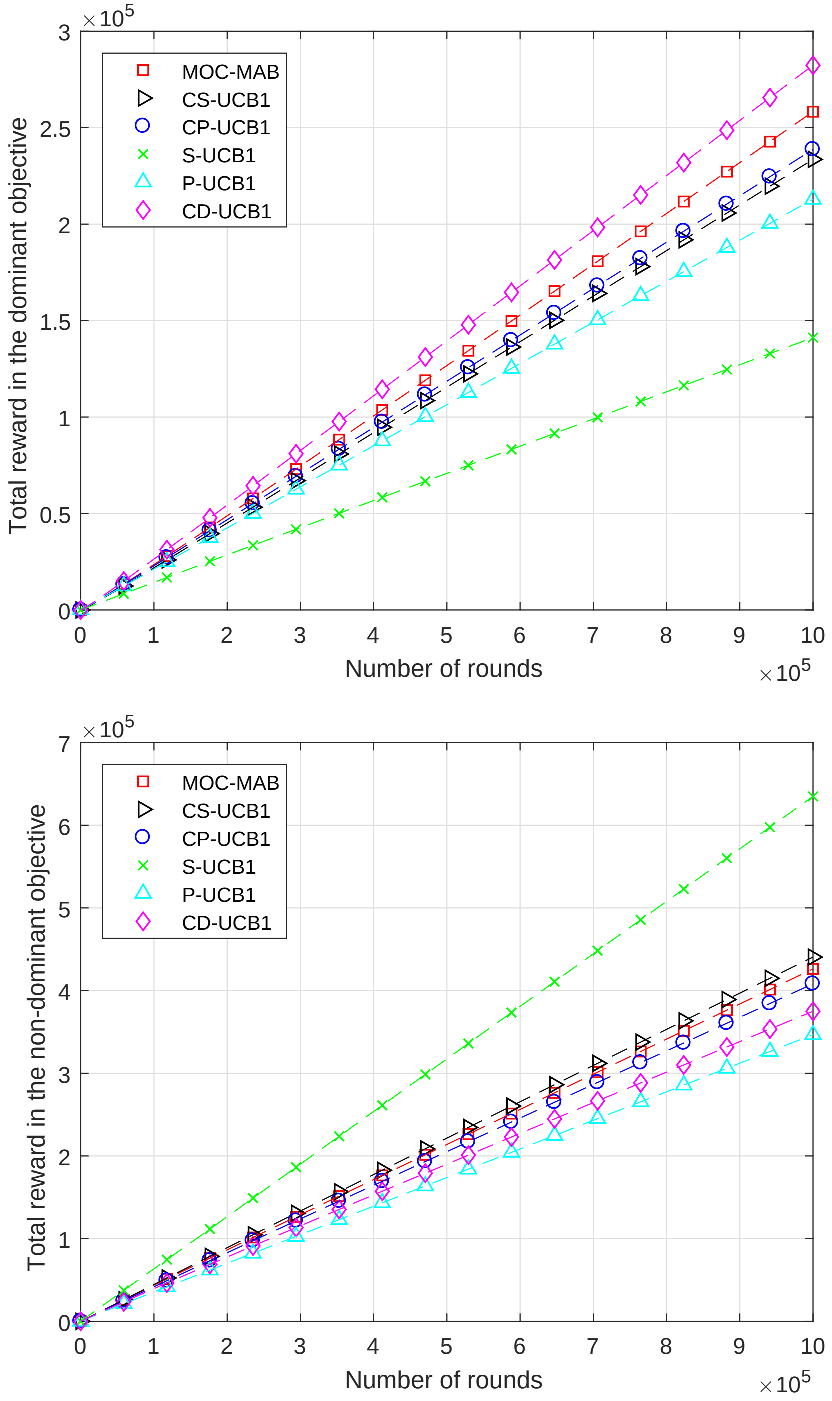

Figure 3.2: Total rewards of MOC-MAB and the other algorithms for Experiment 2.

second objective is related to the average payment. Essentially, when action a is taken in round t, then $r_t^2 = 0$, and $r_t^1 = 0$ if the user does not 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,[9] 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 [46]. 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.3.3 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,

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

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.

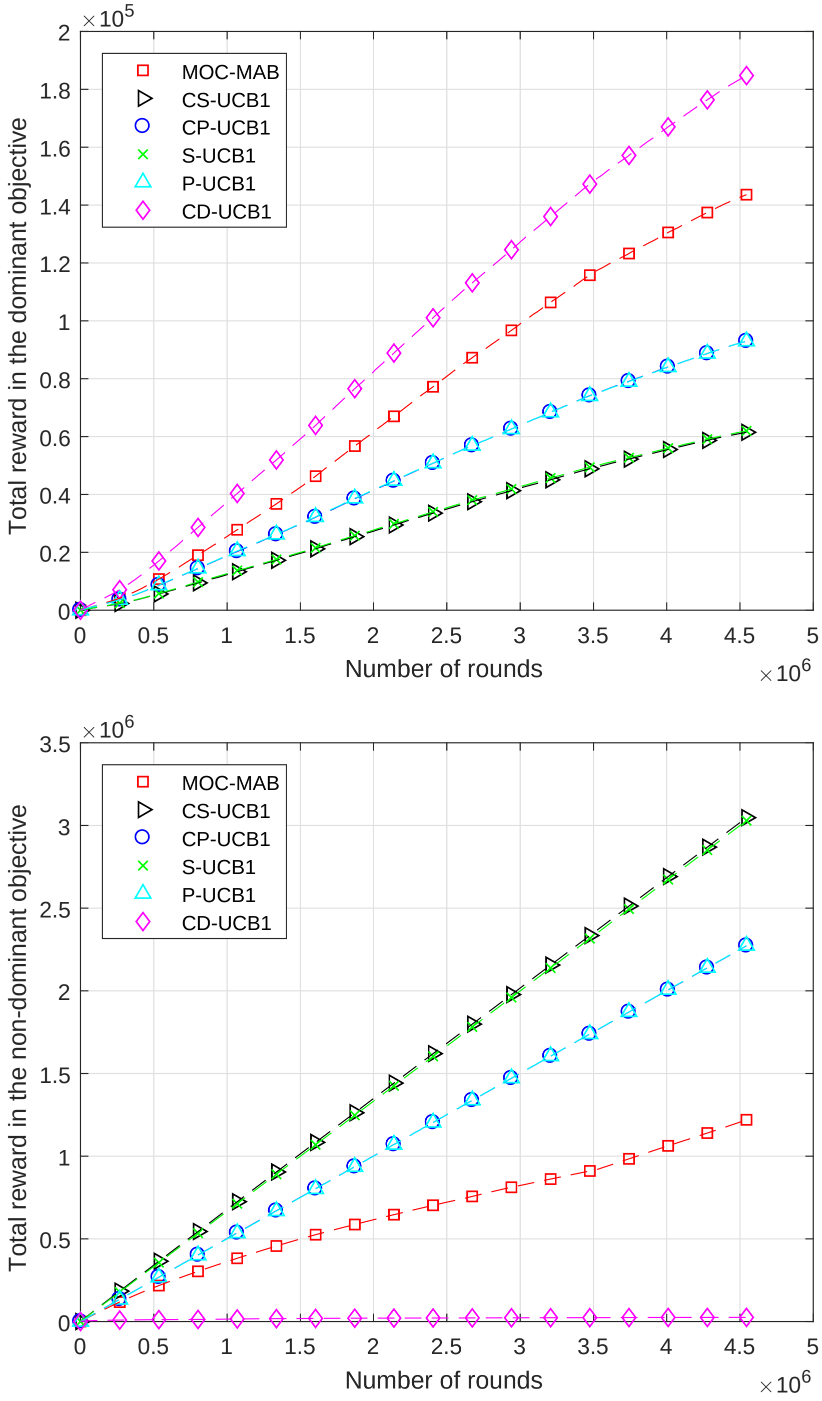

Figure 3.3: Total rewards of MOC-MAB and the other algorithms for Experiment 3.

Chapter 4

Multi-objective Contextual $\mathcal{X}$-armed Bandit

In this chapter, we consider a multi-objective contextual MAB problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set (also called multi-objective contextual $\mathcal{X}$-armed bandit problem). First section includes problem formulation and the definition of the contextual Pareto regret (referred to as the Pareto regret hereafter), which is defined in [19] for two objectives. Here, we extend it to work for an arbitrary number of objectives. For this problem, we propose an online learning algorithm called Pareto Contextual Zooming (PCZ) in Section 4.2 and its Pareto regret is analyzed in Section 4.3. We prove that the regret bound of PCZ is nearly optimal by providing an almost matching lower bound in Section 4.4. Lastly, we evaluate the performance of PCZ in synthetic dataset in Section 4.5. This work was published in [12].[1]

[1] E. Turgay, D. Oner, and C. Tekin, ”Multi-objective contextual bandit problem with similarity information” in *Proc. 21st. Int. Conf. on Artificial Intelligence and Statistics*, pp. 1673-1681, 2018.

4.1 Problem Formulation

The system operates in rounds indexed by $t \in \{1, 2, \ldots\}$. At the beginning of each round, the learner observes a context x_t that comes from a d_x-dimensional context set $\mathcal{X}$. Then, the learner chooses an arm a_t from a d_a-dimensional arm set $\mathcal{A}$. After choosing the arm, the learner obtains a d_r-dimensional random reward vector $r_t := (r_t^1, \ldots, r_t^{d_r})$ where r_t^i denotes the reward obtained from objective $i \in \{1, \ldots, d_r\}$ in round t. Let $\mu_a^i(x)$ denote the expected reward of arm a in objective i for context x and $\mu_a(x) := (\mu_a^1(x), \ldots, \mu_a^{d_r}(x))$. The random reward vector obtained from arm a_t in round t is given as $r_t := \mu_{a_t}(x_t) + \kappa_t$ where κ_t is the d_r-dimensional noise process whose marginal distribution for each objective is conditionally 1-sub-Gaussian, i.e., $\forall \lambda \in \mathbb{R}$

$$\mathrm{E}[e^{\lambda \kappa_t^i} \mid a_{1:t}, x_{1:t}, \kappa_{1:t-1}] \leq \exp(\lambda^2/2)$$

where $b_{1:t} := (b_1, \ldots, b_t)$. Context and arm sets together constitute the set of feasible context-arm pairs, denoted by $\mathcal{F} := \mathcal{X} \times \mathcal{A}$. We assume that the Lipschitz condition holds for the set of feasible context-arm pairs with respect to the expected rewards for all objectives.

Assumption 3. *For all $i \in \{1, \ldots, d_r\}$, $a, a' \in \mathcal{A}$ and $x, x' \in \mathcal{X}$, we have*

$$|\mu_a^i(x) - \mu_{a'}^i(x')| \leq D((x, a), (x', a'))$$

where D is the distance function known by the learner such that $D((x, a), (x', a')) \leq 1$ for all $x, x' \in \mathcal{X}$ and $a, a' \in \mathcal{A}$.

$(\mathcal{F}, D)$ denotes the *similarity space*. In the multi-objective bandit problem, since the objectives might be conflicting, finding an arm that simultaneously maximizes the expected reward in all objectives is in general not possible. Thus, an intuitive way to define optimality is to use the notion of Pareto optimality.

Definition 3 (Pareto optimality)**.** *(i) An arm a is* weakly dominated *by arm a' given context x, denoted by $\mu_a(x) \preceq \mu_{a'}(x)$ or $\mu_{a'}(x) \succeq \mu_a(x)$, if $\mu_a^i(x) \leq \mu_{a'}^i(x), \forall i \in \{1, \ldots, d_r\}$.*

(ii) An arm a is dominated *by arm a' given context x, denoted by $\mu_a(x) \prec \mu_{a'}(x)$ or $\mu_{a'}(x) \succ \mu_a(x)$, if it is weakly dominated and $\exists i \in \{1, \ldots, d_r\}$ such that $\mu_a^i(x) < \mu_{a'}^i(x)$.*
(iii) Two arms a and a' are incomparable given context x, denoted by $\mu_a(x)||\mu_{a'}(x)$, 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. The set of all Pareto optimal arms given a particular context x, is called the* Pareto front, *and is denoted by $\mathcal{O}(x)$.*

The expected loss incurred by the learner in a round due to not choosing an arm in the Pareto front is equal to the Pareto suboptimality gap (PSG) of the chosen arm, which is defined as follows.

Definition 4 (PSG)**.** *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 $\mu_a(x)$ such that a becomes a member of the Pareto front. Formally,*

$$\Delta_a(x) := \inf_{\epsilon \geq 0} \epsilon \quad s.t. \quad (\mu_a(x) + \boldsymbol{\epsilon}) \,||\, \mu_{a'}(x), \forall a' \in \mathcal{O}(x)$$

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

We evaluate the performance of the learner using the Pareto regret, which is given as

$$\mathrm{PR}(T) := \sum_{t=1}^{T} \Delta_{a_t}(x_t). \tag{4.1}$$

The Pareto regret measures the total loss due to playing arms that are not in the Pareto front. The goal of the learner is to minimize its Pareto regret while ensuring fairness over the Pareto optimal arms for the observed contexts. Our regret bounds depend on the Pareto zooming dimension, which is defined below.

Definition 5. *(i) $\hat{\mathcal{F}} \subset \mathcal{F}$ is called an r-packing of $\mathcal{F}$ if all $z, z' \in \hat{\mathcal{F}}$ satisfies $D(z, z') \geq r$. For any $r > 0$, the r-packing number of $\mathcal{F}$ is $A_r^{packing}(\mathcal{F}) := \max\{|\hat{\mathcal{F}}| : \hat{\mathcal{F}}$ is an r-packing of $\mathcal{F}\}$.*
(ii) For a given $r > 0$, let $\mathcal{F}_{\mu,r} := \{(x, a) \in \mathcal{F} : \Delta_a(x) \leq 12r\}$ denote the set of

near-optimal context-arm pairs. The Pareto r-zooming number N_r is defined as the r-packing number of $\mathcal{F}_{\mu,r}$.
(iii) The Pareto zooming dimension given any constant $\tilde{c} > 0$ is defined as $d_p(\tilde{c}) := \inf\{d > 0 : N_r \leq \tilde{c}r^{-d}, \forall r \in (0,1)\}$. With an abuse of notation we let $d_p := d_p(p)$ for $p > 0$.

4.2 Pareto Contextual Zooming Algorithm (PCZ)

In this section, we present a multi-objective contextual bandit algorithm called *Pareto Contextual Zooming* (PCZ). Pseudo-code of PCZ is given in Algorithm 2. PCZ is a multi-objective extension of the single objective contextual zooming algorithm [7]. The proposed algorithm partitions the similarity space non-uniformly according to the arms selected, and the contexts and rewards observed in the past by using a set of *active balls* $\mathcal{B}$ which may change from round to round. Each active ball $B \in \mathcal{B}$ has a radius $r(B)$, center (x_B, a_B) and a domain in the similarity space. The domain of ball B at the beginning of round t is denoted by $\mathrm{dom}_t(B)$, and is defined as the subset of B that excludes all active balls at the beginning of round t that have radius strictly smaller than $r(B)$, i.e., $\mathrm{dom}_t(B) := B \setminus (\cup_{B' \in \mathcal{B}: r(B') < r(B)} B')$. The domains of all active balls cover the similarity space.

Initially, PCZ takes as inputs the time horizon T,[2] the similarity space $(\mathcal{F}, D)$, the confidence parameter $\delta \in (0,1)$, and creates an active ball centered at a random point in the similarity space with radius 1 whose domain covers the entire similarity space. At the beginning of round t, PCZ observes the context x_t, and finds the set of relevant balls denoted by $\hat{\mathcal{R}}(x_t) := \{B \in \mathcal{B} : (x_t, a) \in \mathrm{dom}_t(B)$ for some $a \in \mathcal{A}\}$.

After finding the set of relevant balls, PCZ uses the principle of *optimism under*

[2] While PCZ requires T as input, it is straightforward to extend it to work without knowing T beforehand, by using a standard method, called the doubling trick.

the face of uncertainty to select a ball and an arm. In this principle, the estimated rewards of the balls are inflated by a certain level, such that the inflated reward estimates (also called indices) become an upper confidence bound (UCB) for the expected reward with high probability. Then, PCZ selects a ball whose index is in the Pareto front. In general, this allows the balls that are rarely selected to get explored (because their indices will remain high due to the sample uncertainty being high), which enables the learner to discover new balls that are potentially better than the frequently selected balls in terms of the rewards.

Algorithm 2 Pareto Contextual Zooming

```
Input: $(\mathcal{F}, D)$, $T$, $\delta$
Data: Collection $\mathcal{B}$ of ”active balls” in $(\mathcal{F}, \mathcal{D})$; counters $N_B$ and estimates
    $\hat{\mu}^i_B$, $\forall B \in \mathcal{B}$, $\forall i \in \{1, \dots, d_r\}$
Init: Create ball $B$, with $r(B) = 1$ and an arbitrary center in $\mathcal{F}$. $\mathcal{B} \leftarrow \{B\}$
$\hat{\mu}^i_B = 0$, $\forall i \in \{1, \dots, d_r\}$ and $N_B = 0$
while $1 \leq t \leq T$ do
    Observe $x_t$
    $\hat{\mathcal{R}}(x_t) \leftarrow \{B \in \mathcal{B} : (x_t, a) \in \mathrm{dom}_t(B)$ for some $a \in \mathcal{A}\}$
    $\hat{\mathcal{B}}^* \leftarrow \{B \in \hat{\mathcal{R}}(x_t) : g_B \not\prec g_{B'}, \forall B' \in \hat{\mathcal{R}}(x_t)\}$
    Select an arm $a_t$ uniformly at random from $\{a : (x_t, a) \in \cup_{B \in \hat{\mathcal{B}}^*} \mathrm{dom}_t(B)\}$,
    and observe the rewards $r^i$, $\forall i \in \{1, \dots, d_r\}$
    Uniformly at random choose a ball $\hat{B} \in \hat{\mathcal{B}}^*$ whose domain contains $(x_t, a_t)$
    if $u_{\hat{B}} \leq r(\hat{B})$ then
        Activate (create) a new ball $B'$ whose center is $(x_t, a_t)$ and radius is
        $r(B') = r(\hat{B})/2$
        $\mathcal{B} \leftarrow \mathcal{B} \cup B'$, and $\hat{\mu}^i_{B'} = N_{B'} = 0$, $\forall i \in \{1, \dots, d_r\}$.
        Update the domains of balls in $\mathcal{B}$.
    end if
    Update estimates $\hat{\mu}^i_{\hat{B}} = ((\hat{\mu}^i_{\hat{B}} N_{\hat{B}}) + r^i)/(N_{\hat{B}} + 1)$, $\forall i \in \{1, \dots, d_r\}$ and the
    counter $N_{\hat{B}} = N_{\hat{B}} + 1$
end while
```

The index for each relevant ball is calculated in the following way: First, PCZ computes a pre-index for each ball in $\mathcal{B}$ given by

$$g_B^{i,pre} := \hat{\mu}^i_B + u_B + r(B), \ i \in \{1, \dots, d_r\}$$

which is the sum of the sample mean reward $\hat{\mu}^i_B$, the sample uncertainty $u_B := \sqrt{2A_B/N_B}$, where $A_B := (1 + 2\log(2\sqrt{2} d_r T^{\frac{3}{2}}/\delta))$ and N_B is the number of times ball B is chosen, and the contextual uncertainty $r(B)$. The sample uncertainty represents the uncertainty in the sample mean estimate of the expected reward due to the limited number of random samples in ball B that are used to form this estimate. On the other hand, the contextual uncertainty represents the uncertainty in the sample mean reward due to the dissimilarity of the contexts that lie within ball B. When a ball is selected, its sample uncertainty decreases but its contextual uncertainty is always fixed. Essentially, the pre-index $g_B^{i,pre}$ is a UCB for the expected reward at the center of ball B in objective i. PCZ uses these pre-indices to compute an index for each relevant ball, given as

$$\begin{aligned} g_B &:= (g_B^1, \ldots, g_B^{d_r}) \text{ where} \\ g_B^i &:= r(B) + \min_{B' \in \mathcal{B}} (g_{B'}^{i,pre} + D(B', B)) \end{aligned}$$

and $D(B', B)$ represents the distance between the centers of the balls B' and B in the similarity space.

After the indices of the relevant balls are calculated, PCZ computes the Pareto front among the set of balls in $\hat{\mathcal{R}}(x_t)$ by using g_B for ball $B \in \hat{\mathcal{R}}(x_t)$ as a proxy for the expected reward (see Definition 3), which is given as $\hat{\mathcal{B}}^* := \{B \in \hat{\mathcal{R}}(x_t) : g_B \not\prec g_{B'}, \forall B' \in \hat{\mathcal{R}}(x_t)\}$, and uniformly at random selects an arm in the union of the domains of the balls whose indices are in the Pareto front. After observing the reward of the selected arm, it uniformly at random picks a ball whose index is in the Pareto front and contains (x_t, a_t), updates its parameters, and the above procedure repeats in the next round. This selection ensures fairness over the estimated set of Pareto optimal arms in each round.

The remaining important issue is how to create the balls in order to trade-off sample uncertainty and contextual uncertainty in an optimal way. This is a difficult problem due to the fact that the learner does not know how contexts arrive beforehand, and the rate that the contexts arrive in different regions of the context set can dynamically change over time. To overcome these issues, PCZ uses the concept of contextual zooming. Basically, PCZ adaptively partitions the context space according to the past context arrivals, arm selections and obtained rewards. Specifically, the radius of the balls are adjusted to optimize the two sources of error described above. For this, when the sample uncertainty of the selected ball B is found to be smaller than or equal to the radius of the ball (say in round t), a new child ball B' centered at (x_t, a_t) whose radius is equal to the half of the parent ball's (ball B's) radius is created, and the domains of all active balls are updated.[3] Note that when a child ball is created, it does not contain any sample, so its sample uncertainty is infinite. This results in an infinite pre-index. For this reason, g^i_B is used instead of $g^{i,pre}_B$ as an enhanced UCB, which is in general tighter than the UCB based on the pre-index.

4.3 Regret Analysis of PCZ

In this section, we prove that PCZ achieves (i) $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ Pareto regret with probability at least $1-\delta$, and (ii) $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ expected Pareto regret, where d_p is the Pareto zooming dimension.

First, we define the variables that will be used in the Pareto regret analysis. For an event $\mathcal{H}$, let $\mathcal{H}^c$ denote the complement of that event. For all the parameters defined in the pseudo-code of PCZ, we explicitly use the round index t, when referring to the value of that parameter at the beginning of round t. For instance, $N_B(t)$ denotes the value of N_B at the beginning of round t, and $\mathcal{B}(t)$ denotes the set of balls created by PCZ by the beginning of round t. Let $\mathcal{B}'(T)$ denote the set of balls chosen at least once by the end of round T. Note that $\mathcal{B}'(T) \subset \mathcal{B}(T)$ and $\mathcal{B}(T)$ is a random variable that depends both on the contexts arrivals, selected arms and observed rewards. Let $R^i_B(t)$ denote the random reward of ball B in objective i at round t and let $\tau_B(t)$ denote the first round after the round in which ball B is chosen by PCZ for the tth time. Moreover, the round that comes just after $B \in \mathcal{B}(T)$ is created is denoted by $\tau_B(0)$, and the domain of B when it is created is denoted by $\text{dom}(B)$. Hence, $\text{dom}_t(B) \subseteq \text{dom}(B)$, $\forall t \in \{\tau_B(0), \ldots, T\}$. For $B \in \mathcal{B}'(T)$, let $\mathcal{T}_B$ denote the set of rounds in $\{\tau_B(0), \ldots, T\}$ in which ball B is selected by PCZ. Also let $\tilde{x}_B(t) := x_{\tau_B(t)-1}$, $\tilde{a}_B(t) := a_{\tau_B(t)-1}$, $\tilde{R}^i_B(t) := R^i_B(\tau_B(t)-1)$, $\tilde{\kappa}^i_B(t) := \kappa^i_{\tau_B(t)-1}$, $\tilde{N}_B(t) := N_B(\tau_B(t))$, $\tilde{\mu}^i_B(t) := \hat{\mu}^i_B(\tau_B(t))$, $\tilde{g}^{i,pre}_B(t) := g^{i,pre}_B(\tau_B(t))$, $\tilde{g}^i_B(t) := g^i_B(\tau_B(t))$, and $\tilde{u}_B(t) := u_B(\tau_B(t))$. We note

[3] We also use B^{par} to denote the parent of a ball B.

that all inequalities that involve random variables hold with probability one unless otherwise stated. Let

$$\mathrm{PR}_B(T) := \sum_{t=1}^{N_B(T+1)} \Delta_{\tilde{a}_B(t)}(\tilde{x}_B(t))$$

denote the Pareto regret incurred in ball $B \in \mathcal{B}'(T)$ for rounds in $\mathcal{T}_B$. Then, the Pareto regret in (4.1) can be written as

$$\mathrm{PR}(T) = \sum_{B \in \mathcal{B}'(T)} \mathrm{PR}_B(T).$$

Next, we define the following lower and upper bounds: $L_B^i(t) := \hat{\mu}_B^i(t) - u_B(t)$ and $U_B^i(t) := \hat{\mu}_B^i(t) + u_B(t)$, $\tilde{L}_B^i(t) := \tilde{\mu}_B^i(t) - \tilde{u}_B(t)$ and $\tilde{U}_B^i(t) := \tilde{\mu}_B^i(t) + \tilde{u}_B(t)$ for $i \in \{1, \ldots, d_r\}$. Let

$$\mathrm{UC}_B^i := \bigcup_{t=\tau_B(0)}^{T+1} \{\mu_{a_B}^i(x_B) \notin [L_B^i(t) - r(B), U_B^i(t) + r(B)]\}$$

denote the event that the learner is not confident about its reward estimate in objective i in ball B for at least once from round $\tau_B(0)$ to round T, and

$$\tilde{\mathrm{UC}}_B^i := \bigcup_{t=0}^{N_B(T+1)} \{\mu_{a_B}^i(x_B) \notin [\tilde{L}_B^i(t) - r(B), \tilde{U}_B^i(t) + r(B)]\}.$$

Let $\tau_B(N_B(T+1)+1) = T+2$. Then, for $z = 0, \ldots, N_B(T+1)$ the events $\{\mu_{a_B}^i(x_B) \notin [L_B^i(t) - r(B), U_B^i(t) + r(B)]\}$ and $\{\mu_{a_B}^i(x_B) \notin [L_B^i(t') - r(B), U_B^i(t') + r(B)]\}$ are identical for any $t, t' \in \{\tau_B(z), \ldots, \tau_B(z+1) - 1\}$. Thus,

$$\begin{aligned}&\bigcup_{t=\tau_B(z)}^{\tau_B(z+1)-1} \{\mu_{a_B}^i(x_B) \notin [L_B^i(t) - r(B), U_B^i(t) + r(B)]\} \\ &= \{\mu_{a_B}^i(x_B) \notin [\tilde{L}_B^i(z) - r(B), \tilde{U}_B^i(z) + r(B)]\}.\end{aligned}$$

Hence, we conclude that $\mathrm{UC}_B^i = \tilde{\mathrm{UC}}_B^i$. Let $\mathrm{UC}_B := \cup_{i \in \{1,\ldots,d_r\}} \mathrm{UC}_B^i$ and $\mathrm{UC} := \cup_{B \in \mathcal{B}(T)} \mathrm{UC}_B$. Next, we will bound $\mathrm{E}[\mathrm{PR}(T)]$. We have

$$\begin{aligned}\mathrm{E}[\mathrm{PR}(T)] &= \mathrm{E}[\mathrm{PR}(T) \mid \mathrm{UC}] \Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{PR}(T) \mid \mathrm{UC}^c] \Pr(\mathrm{UC}^c) \\ &\leq C_{\max} T \Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{PR}(T) \mid \mathrm{UC}^c] \qquad (4.2)\end{aligned}$$

where $C_{\max} := \sup_{(x,a)\in\mathcal{F}} \Delta_a(x)$. Since $\mathcal{B}(T)$ is a random variable, we have

$$\Pr(\text{UC}) = \int \Pr(\text{UC}|\mathcal{B}(T)) dQ(\mathcal{B}(T)) \tag{4.3}$$

where $Q(\mathcal{B}(T))$ denotes the distribution of $\mathcal{B}(T)$. We also have

$$\Pr(\text{UC}|\mathcal{B}(T)) = \Pr(\cup_{B\in\mathcal{B}(T)} \text{UC}_B|\mathcal{B}(T)) \leq \sum_{B\in\mathcal{B}(T)} \Pr(\text{UC}_B|\mathcal{B}(T)). \tag{4.4}$$

We proceed by bounding the term Pr(UC) in (4.2). For this, first we bound $\Pr(\text{UC}_B|\mathcal{B}(T))$ in the next lemma.

Lemma 7. *When PCZ is run, we have* $\Pr(UC_B|\mathcal{B}(T)) \leq \delta/T$, $\forall B \in \mathcal{B}(T)$.

Proof. From the definitions of $\tilde{L}^i_B(t)$, $\tilde{U}^i_B(t)$ and $\tilde{\text{UC}}^i_B$, it can be observed that the event UC^i_B happens when $\tilde{\mu}^i_B(t)$ remains away from $\mu^i_{a_B}(x_B)$ for some $t \in \{0, \ldots, N_B(T+1)\}$. Using this information, we can use the concentration inequality given in Appendix B. In this formulation expected rewards of the arms must be equal in all rounds, but in our case, $\mu^i_{\tilde{a}_B(t)}(\tilde{x}_B(t))$ changes since the elements of $\{\tilde{x}_B(t), \tilde{a}_B(t)\}_{t=1}^{N_B(T+1)}$ are not identical which makes distributions of $\tilde{R}^i_B(t)$, $t \in \{1, \ldots, N_B(T+1)\}$ different.

In order to overcome this issue, we use the sandwich technique proposed in [6] and later used in [19]. For any ball $B \in \mathcal{B}(T)$, we have $\Pr(\mu^i_{a_B}(x_B) \notin [\tilde{L}^i_B(0) - r(B), \tilde{U}^i_B(0) + r(B)]) = 0$ since $\tilde{\mu}^i_B(0) = 0$, $\tilde{L}^i_B(0) = -\infty$ and $\tilde{U}^i_B(0) = \infty$. Thus, for $B \in \mathcal{B}(T) \setminus \mathcal{B}'(T)$, we have $\Pr(\text{UC}_B|\mathcal{B}(T)) = 0$. Hence, we proceed by bounding the probabilities of the events $\{\mu^i_{a_B}(x_B) \notin [\tilde{L}^i_B(t) - r(B), \tilde{U}^i_B(t) + r(B)]\}$, for $t > 0$ and for the balls in $\mathcal{B}'(T)$. Recall that $\tilde{R}^i_B(t) = \mu^i_{\tilde{a}_B(t)}(\tilde{x}_B(t)) + \tilde{\kappa}^i_B(t)$ and $\tilde{\mu}^i_B(t) = \sum_{l=1}^{t} \tilde{R}^i_B(l)/t$ (for $t > 0$ and $B \in \mathcal{B}'(T)$). For each $i \in \{1, \ldots, d_r\}$, $B \in \mathcal{B}'(T)$, let

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

We define two new sequences of random variables, whose sample mean values will lower and upper bound $\tilde{\mu}^i_B(t)$. The *best sequence* is defined as $\{\bar{R}^i_B(t)\}_{t=1}^{N_B(T+1)}$ where $\overline{R}^i_B(t) := \overline{\mu}^i_B + \tilde{\kappa}^i_B(t)$, and the *worst sequence* is defined as $\{\underline{R}^i_B(t)\}_{t=1}^{N_B(T+1)}$ where $\underline{R}^i_B(t) := \underline{\mu}^i_B + \tilde{\kappa}^i_B(t)$. Let $\overline{\mu}^i_B(t) := \sum_{l=1}^t \overline{R}^i_B(l)/t$ and $\underline{\mu}^i_B(t) := \sum_{l=1}^t \underline{R}^i_B(l)/t$. We have

$$\underline{\mu}^i_B(t) \leq \tilde{\mu}^i_B(t) \leq \overline{\mu}^i_B(t) \;\; \forall t \in \{1, \ldots, N_B(T+1)\}.$$

Let

$$\begin{aligned}
\overline{L}^i_B(t) &:= \overline{\mu}^i_B(t) - \tilde{u}_B(t) \\
\overline{U}^i_B(t) &:= \overline{\mu}^i_B(t) + \tilde{u}_B(t) \\
\underline{L}^i_B(t) &:= \underline{\mu}^i_B(t) - \tilde{u}_B(t) \\
\underline{U}^i_B(t) &:= \underline{\mu}^i_B(t) + \tilde{u}_B(t).
\end{aligned}$$

It can be shown that

$$\begin{aligned}
&\{\mu^i_{a_B}(x_B) \notin [\tilde{L}^i_B(t) - r(B), \tilde{U}^i_B(t) + r(B)]\} \qquad (4.5)\\
&\subset \{\mu^i_{a_B}(x_B) \notin [\overline{L}^i_B(t) - r(B), \overline{U}^i_B(t) + r(B)]\} \\
&\cup \{\mu^i_{a_B}(x_B) \notin [\underline{L}^i_B(t) - r(B), \underline{U}^i_B(t) + r(B)]\}.
\end{aligned}$$

Moreover, the following inequalities can be obtained from Assumption 3:

$$\mu^i_{a_B}(x_B) \leq \overline{\mu}^i_B \leq \mu^i_{a_B}(x_B) + r(B) \qquad (4.6)$$

$$\mu^i_{a_B}(x_B) - r(B) \leq \underline{\mu}^i_B \leq \mu^i_{a_B}(x_B). \qquad (4.7)$$

Using (4.6) and (4.7) it can be shown that

$$\begin{aligned}
&\{\mu^i_{a_B}(x_B) \notin [\overline{L}^i_B(t) - r(B), \overline{U}^i_B(t) + r(B)]\} \subset \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\}, \\
&\{\mu^i_{a_B}(x_B) \notin [\underline{L}^i_B(t) - r(B), \underline{U}^i_B(t) + r(B)]\} \subset \{\underline{\mu}^i_B \notin [\underline{L}^i_B(t), \underline{U}^i_B(t)]\}.
\end{aligned}$$

Plugging this to (4.5), we get

$$\begin{aligned}
&\{\mu^i_{a_B}(x_B) \notin [\tilde{L}^i_B(t) - r(B), \tilde{U}^i_B(t) + r(B)]\} \\
&\subset \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\} \bigcup \{\underline{\mu}^i_B \notin [\underline{L}^i_B(t), \underline{U}^i_B(t)]\}.
\end{aligned}$$

Using the equation above and the union bound we obtain

$$\Pr(\mathrm{UC}^i_B | \mathcal{B}(T)) \leq \Pr\left(\bigcup_{t=1}^{N_B(T+1)} \{\overline{\mu}^i_B \notin [\overline{L}^i_B(t), \overline{U}^i_B(t)]\} \right)$$

$$+ \Pr\left(\bigcup_{t=1}^{N_B(T+1)} \{\underline{\mu}_B^i \notin [\underline{L}_B^i(t), \underline{U}_B^i(t)]\} \right).$$

Both terms on the right-hand side of the inequality above can be bounded using the concentration inequality in Appendix B. Using $\theta = \delta/(2d_r T)$, in Appendix B gives $\Pr(\mathrm{UC}_B^i|\mathcal{B}(T)) \leq \delta/(d_r T)$, since $1 + N_B(t) \leq 1 + N_B(T+1) \leq 2T$. Then, using the union bound over all objectives, we obtain $\Pr(\mathrm{UC}_B|\mathcal{B}(T)) \leq \delta/T$. □

Next, we bound Pr(UC). Since $|\mathcal{B}(T)| \leq T$, by using union bound over all created balls, (4.3) and (4.4) we obtain

$$\begin{aligned} \Pr(\mathrm{UC}) &\leq \int \Pr(\mathrm{UC}|\mathcal{B}(T)) dQ(\mathcal{B}(T)) \\ &\leq \int \Bigg(\sum_{B \in \mathcal{B}(T)} \Pr(\mathrm{UC}_B|\mathcal{B}(T)) \Bigg) dQ(\mathcal{B}(T)) \\ &\leq \frac{\delta T}{T} \int dQ(\mathcal{B}(T)) \leq \delta. \end{aligned}$$

Then, by using (4.2),

$$\mathrm{E}[\mathrm{PR}(T)] \leq C_{\max} T \Pr(\mathrm{UC}) + \mathrm{E}[\mathrm{PR}(T) \mid \mathrm{UC}^c] \leq C_{\max} T \delta + \mathrm{E}[\mathrm{PR}(T) \mid \mathrm{UC}^c]. \tag{4.8}$$

In the remainder of the analysis we bound $\mathrm{PR}(T)$ on event UC^c. Then, we conclude the analysis by using this to bound 4.8. For simplicity of notation, in the following lemmas we use $B(t)$ to denote the ball selected in round t.

Lemma 8. *Consider a virtual arm with expected reward vector $g_{B(t)}(t)$, where $g^i_{B(t)}(t)$ denotes the expected reward in objective i. On event UC^c, we have*

$$g_{B(t)}(t) \not\prec \mu_a(x_t), \forall a \in \mathcal{A}.$$

Proof. For any relevant ball $B \in \hat{\mathcal{R}}(x_t)$ and $i \in \{1, \ldots, d_r\}$, let $\tilde{B}_i := \arg\min_{B' \in \mathcal{B}(t)} \{g^{i,pre}_{B'}(t) + D(B, B')\}$. Then, $\forall a$ such that $(x_t, a) \in \mathrm{dom}_t(B)$, we have

$$g^i_B(t) = r(B) + g^{i,pre}_{\tilde{B}_i}(t) + D(B, \tilde{B}_i)$$

$$
\begin{aligned}
&= r(B) + \hat{\mu}^i_{\tilde{B}_i}(t) + u_{\tilde{B}_i}(t) + r(\tilde{B}_i) + D(B, \tilde{B}_i) \\
&= r(B) + U^i_{\tilde{B}_i}(t) + r(\tilde{B}_i) + D(B, \tilde{B}_i) \\
&\geq r(B) + \mu^i_{a_{\tilde{B}_i}}(x_{\tilde{B}_i}) + D(B, \tilde{B}_i) \\
&\geq r(B) + \mu^i_{a_B}(x_B) \geq \mu^i_a(x_t)
\end{aligned}
$$

where the first inequality holds by the definition of UC^c, and the second and third inequalities hold by Assumption 3. According to the above inequality $g_B(t) \succeq \mu_a(x_t)$, $\forall a$ such that $(x_t, a) \in \mathrm{dom}_t(B)$. We also know that $\forall a \in \mathcal{A}$, $\exists B \in \hat{\mathcal{R}}(x_t)$ such that $(x_t, a) \in \mathrm{dom}_t(B)$. Moreover, by the selection rule of PCZ, $g_{B(t)}(t) \not\prec g_B(t)$ for all $B \in \hat{\mathcal{R}}(x_t)$. By combining these results, $g_{B(t)}(t) \not\prec \mu_a(x_t), \forall a \in \mathcal{A}$, and hence, the virtual arm with expected reward vector $g_{B(t)}(t)$ is not dominated by any of the arms. □

Lemma 9. *When PCZ is run, on event UC^c, we have*

$$\Delta_{a_t}(x_t) \leq 14r(B(t)) \quad \forall t \in \{1, \ldots, T\}.$$

Proof. This proof is similar to the proof in [7]. To bound the PSG of the selected ball, we first bound the index of the selected ball $g^i_{B(t)}(t)$. Recall that $B^{par}(t)$ denotes the parent ball of the selected ball $B(t)$.[4] We have

$$
\begin{aligned}
g^{i,pre}_{B^{par}(t)}(t) &= \hat{\mu}^i_{B^{par}(t)}(t) + r(B^{par}(t)) + u_{B^{par}(t)}(t) \\
&= L^i_{B^{par}(t)}(t) + r(B^{par}(t)) + 2u_{B^{par}(t)}(t) \\
&\leq \mu^i_{a_{B^{par}(t)}}(x_{B^{par}(t)}) + 2r(B^{par}(t)) + 2u_{B^{par}(t)}(t) \\
&\leq \mu^i_{a_{B^{par}(t)}}(x_{B^{par}(t)}) + 4r(B^{par}(t)) \\
&\leq \mu^i_{a_{B(t)}}(x_{B(t)}) + 5r(B^{par}(t)) \qquad (4.9)
\end{aligned}
$$

where the first inequality holds by the definition of UC^c, the second inequality holds since $u_{B^{par}(t)}(t) \leq r(B^{par}(t))$, and the third inequality holds due to Assumption 3. We also have

$$
\begin{aligned}
g^i_{B(t)}(t) &\leq r(B(t)) + g^{i,pre}_{B^{par}(t)}(t) + D(B^{par}(t), B(t)) \\
&\leq r(B(t)) + g^{i,pre}_{B^{par}(t)}(t) + r(B^{par}(t))
\end{aligned}
$$

[4] The bound for $B(1)$ is trivial since it contains the entire similarity space.

$$\begin{aligned}
&\le r(B(t)) + \mu^i_{a_{B(t)}}(x_{B(t)}) + 6r(B^{par}(t)) \\
&\le \mu^i_{a_{B(t)}}(x_{B(t)}) + 13r(B(t)) \\
&\le \mu^i_{a_t}(x_t) + 14r(B(t))
\end{aligned}$$

where the third inequality follows from (4.9). Since $g^i_{B(t)}(t) - \mu^i_{a_t}(x_t) \le 14r(B(t))$ for all $i \in \{1, \dots, d_r\}$ and the virtual arm is not dominated by any arm in the Pareto front by Lemma 8, the PSG of the selected arm is bounded by $\Delta_{a_t}(x_t) \le 14r(B(t))$. □

Lemma 10. *When PCZ is run, on event UC^c, the maximum number of radius r balls that are created by round T is bounded by the Pareto r-zooming number N_r given in Definition 5. Moreover, in any round t in which a radius r ball is created, we have $\Delta_{a_t}(x_t) \le 12r$.*

Proof. Assume that a new ball is created at round t whose parent is $B(t)$. Let $B'(t)$ denote the created ball. We have,

$$\begin{aligned}
g^i_{B(t)}(t) &\le r(B(t)) + g^{i,pre}_{B(t)}(t) \\
&= \hat{\mu}^i_{B(t)}(t) + 2r(B(t)) + u_{B(t)}(t) \\
&= L^i_{B(t)}(t) + 2r(B(t)) + 2u_{B(t)}(t) \\
&\le \mu^i_{a_{B(t)}}(x_{B(t)}) + 3r(B(t)) + 2u_{B(t)}(t) \\
&\le \mu^i_{a_{B(t)}}(x_{B(t)}) + 5r(B(t)) \\
&\le \mu^i_{a_{B'(t)}}(x_{B'(t)}) + 6r(B(t)) \\
&\le \mu^i_{a_{B'(t)}}(x_{B'(t)}) + 12r(B'(t))
\end{aligned}$$

where the first inequality follows from the definition of $g^i_{B(t)}(t)$, the second inequality holds by the definition of UC^c, the third inequality holds due to the fact that $B(t)$ is a parent ball, the fourth inequality holds due to Assumption 3, and the last inequality follows from $r(B'(t)) = r(B(t))/2$. Similar to the proof of Lemma 9, since $g^i_{B(t)}(t) - \mu^i_{a_{B'(t)}}(x_{B'(t)}) \le 12r(B'(t))$ for all $i \in \{1, \dots, d_r\}$ and the virtual arm is not dominated by any arm in the Pareto front by Lemma 8, the PSG of point $(x_{B'(t)}, a_{B'(t)}) = (x_t, a_t)$ is bounded by $12r(B'(t))$. This implies that center of the ball created in any round t has a PSG that is at most $12r(B'(t))$.

Thus, the center of $B'(t)$ is in $\mathcal{F}_{\mu,r(B'(t))}$. Next, consider any two balls B and B' with radius r created by PCZ. Based on the ball creation and domain update rules of PCZ, the distance between the centers of these balls must be at least r. As a result, the maximum number of radius r balls created is bounded by the r-packing number of $\mathcal{F}_{\mu,r}$, which is N_r. □

The following lemma bounds the regret of PCZ in terms of N_r by using the results in Lemma 10 and 9.

Lemma 11. *On event UC^c, the Pareto regret of PCZ by round T is bounded by $PR(T) \leq 28Tr_0 + \sum_{r=2^{-i}:i\in\mathbb{N},r_0\leq r\leq 1} 56r^{-1}N_r \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta)$ for any $r_0 \in (0,1]$.*

Proof. The maximum number of times a radius r ball B can be selected before it becomes a parent ball is upper bounded by $1 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_r T^{\frac{3}{2}}/\delta))$. From the result of Lemma 9, we know that the Pareto regret in each of these rounds is upper bounded by $14r$. Note that after ball B becomes a parent ball, it will create a new radius $r/2$ child ball every time it is selected. From Lemma 10, we know that the Pareto regret in each of these rounds is bounded above by $12(r/2)$. Therefore, we can include the Pareto regret incurred in a round in which a new child ball with radius r is created from a parent ball as a part of the child ball's (total) Pareto regret. Hence, the Pareto regret incurred in a radius r ball is upper bounded by

$$\begin{aligned}
&14r\left(1 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_r T^{\frac{3}{2}}/\delta))\right) + 12r \\
&\leq 14r\left(2 + 2r^{-2}(1 + 2\log(2\sqrt{2}d_r T^{\frac{3}{2}}/\delta))\right) \\
&\leq 56r^{-1}\log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta).
\end{aligned}$$

Let $r_l := 2^{\lceil \log(r_0)/\log(2) \rceil}$. We have $r_0/2 \leq r_l/2 \leq r_0 \leq r_l \leq 2r_0$. The one-round Pareto regret of the balls whose radii are smaller than r_l is bounded by $14r_l$ on event UC^c according to Lemma 9. Also, we know that $14r_l \leq 28r_0$ by the above inequality. Therefore, the Pareto regret due to all balls with radii smaller than r_l by time T is bounded by $28Tr_0$, and the Pareto regret due to all balls with radii

$r = 2^{-i} \geq r_0$ is bounded by $56r^{-1}N_r \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta)$. Thus, summing this up for all possible balls, we obtain the following Pareto regret bound on event UC^c:

$$\mathrm{PR}(T) \leq 28Tr_0 + \sum_{r=2^{-i}: i\in\mathbb{N}, r_0 \leq r \leq 1} 56r^{-1}N_r \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta).$$

□

The following theorem gives a high probability Pareto regret bound for PCZ.

Theorem 4. *For any $p > 0$, the Pareto regret of PCZ by round T is bounded with probability at least $1 - \delta$ (on event UC^c) by*

$$PR(T) \leq (28 + 112p \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta)) T^{\frac{1+d_p}{2+d_p}}$$

where d_p is given in Definition 5.

Proof. Using Definition 5 and the result of Lemma 11 on event UC^c, we have

$$\begin{aligned}\mathrm{PR}(T) &\leq 28Tr_0 + \sum_{r=2^{-i}: i\in\mathbb{N}}^{r_0 \leq r \leq 1} 56r^{-1}pr^{-d_p} \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta) \\ &\leq 28Tr_0 + 56p \log(2\sqrt{2}d_r T^{\frac{3}{2}} e/\delta) \sum_{r=2^{-i}: i\in\mathbb{N}}^{r_0 \leq r \leq 1} r^{-d_p - 1}.\end{aligned}$$

We obtain the final result by setting $r_0 = T^{\frac{-1}{2+d_p}}$. □

Corollary 2. *When PCZ is run with $\delta = 1/T$, then for any $p > 0$, the expected Pareto regret of PCZ by round T is bounded by*

$$\mathrm{E}[PR(T)] \leq (28 + 112p \log(2\sqrt{2}d_r T^{\frac{5}{2}} e)) T^{\frac{1+d_p}{2+d_p}} + C_{\max}.$$

Proof. Theorem 4 is used in (4.8) with $\delta = 1/T$. □

4.4 Lower Bound of PCZ

It is shown in [7] that for the contextual bandit problem with similarity information the regret lower bound is $\Omega(T^{(1+d_z)/(2+d_z)})$, where d_z is the contextual zooming dimension. We use this to give a lower bound on the Pareto regret. Consider an instance of the multi-objective contextual bandit problem where $\mu^i_a(x) = \mu^j_a(x)$, $\forall (x, a) \in \mathcal{F}$ and $\forall i, j \in \{1, \ldots, d_r\}$, and $\kappa^i_t = \kappa^j_t$, $\forall t \in \{1, \ldots, T\}$ and $\forall i, j \in \{1, \ldots, d_r\}$. In this case, the contextual zooming dimension of all objectives are equal (i.e., all d_zs are equal). Moreover, by definition of the Pareto zooming dimension $d_p = d_z$. Therefore, this case is equivalent to the single objective contextual bandit problem. Hence, our regret bound becomes $\tilde{O}(T^{(1+d_z)/(2+d_z)})$ which matches with the lower bound up to a logarithmic factor.

4.5 Illustrative Results

We evaluate the performance of PCZ on a synthetic dataset. We take $\mathcal{X} = [0, 1]$, $\mathcal{A} = [0, 1]$ and generate $\mu^1_a(x)$ and $\mu^2_a(x)$ as shown in Figure 4.1. To generate $\mu^1_a(x)$, we first define a line by equation $8x + 10a = 8$ and let $a_1(x) = (8 - 8x)/10$. For all context arm pairs (x, a), we set $\mu^1_a(x) = \max\{0, (1 - 5|a - a_1(x)|)\}$. Similarly, to generate $\mu^2_a(x)$, we define the line $8x + 10a = 10$ and let $a_2(x) = (10 - 8x)/10$. Then, we set $\mu^2_a(x) = \max\{0, (1 - 5(a_2(x) - a))\}$ for $a \leq a_2(x)$ and $\mu^2_a(x) = \max\{0, (1 - (a - a_2(x))/4)\}$ for $a > a_2(x)$.

Based on the definitions given above, the Pareto optimal arms given context x lie in the interval $[a_1(x), a_2(x)]$. To evaluate the fairness of PCZ, we define six bins that correspond to context-arm pairs in the Pareto front. Given context x, the 1st bin contains all arms in the interval $[a_1(x), a_1(x) + 1/30]$ and the ith bin $i \in \{2, ..., 6\}$ contains all arms in the interval $(a_1(x) + (i - 1)/30, a_1(x) + i/30]$. Simply, first three bins include the Pareto optimal arms whose expected rewards in the first objective are higher than the expected rewards in the second objective and the last three bins include the Pareto optimal arms whose expected rewards in the second objective are higher than the expected rewards in the first objective.

We assume that the reward of arm a in objective i given context x is a Bernoulli random variable with parameter $\mu^i_a(x)$. In addition, at each round t, the context x_t is sampled from the uniform distribution over $\mathcal{X}$.

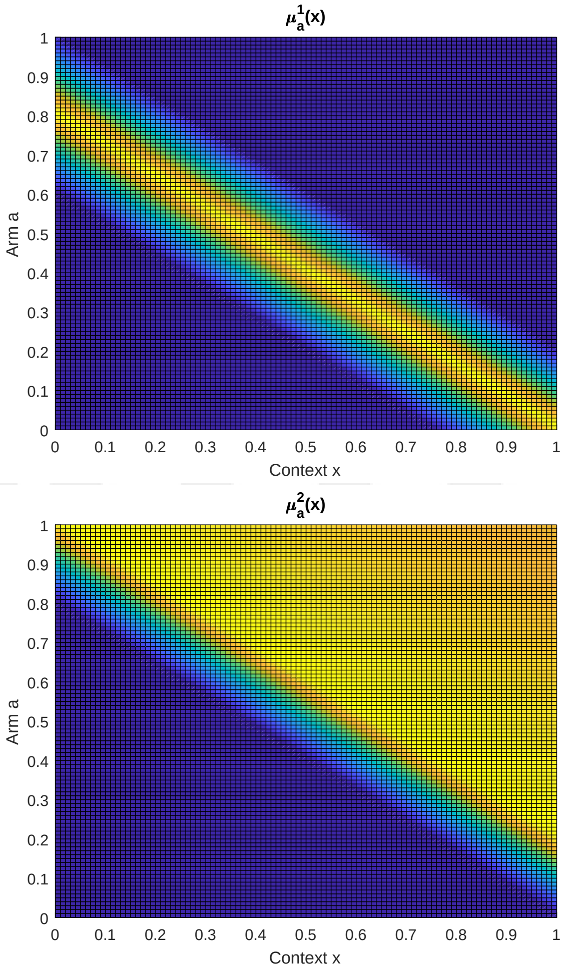

Figure 4.1: Expected reward of Context-Arm Pairs (Yellow represents 1, Dark Blue represents 0)

We compare our algorithm with Contextual Zooming [7] and Random Selection, which chooses in each round an arm uniformly at random from $\mathcal{A}$. Contextual Zooming only uses the rewards in the first objective. Both PCZ and Contextual Zooming uses scaled Euclidean distance.[5] We choose $\delta = 1/T$ in PCZ, set $T = 10^5$, run each algorithm 100 times, and report the average of the results in these runs.

The Pareto Regret is reported in Figure 4.2(i) as a function of the number

[5] We set $D((x,a),(x',a')) = \sqrt{(x-x')^2 + (a-a')^2}/\sqrt{2}$. While this choice does not satisfy Assumption 3, we use this setup to illustrate that learning is still possible when the distance function is not perfectly known by the learner.

of rounds. It is observed that the Pareto regret of PCZ at $T = 10^5$ is 3.61% higher than that of Contextual Zooming and 17.1% smaller than that of Random Selection. We compare the fairness of the algorithms in Figure 4.2(ii). For this, we report the selection ratio of each Pareto front bin, which is defined for bin i as the number of times a context-arm pair in bin i is selected divided by the number of times a Pareto optimal arm is selected by round T. We observe that the selection ratio of all bins are almost the same for PCZ, while Contextual Zooming selects the context-arm pairs in the 1st bin much more than the other bins.

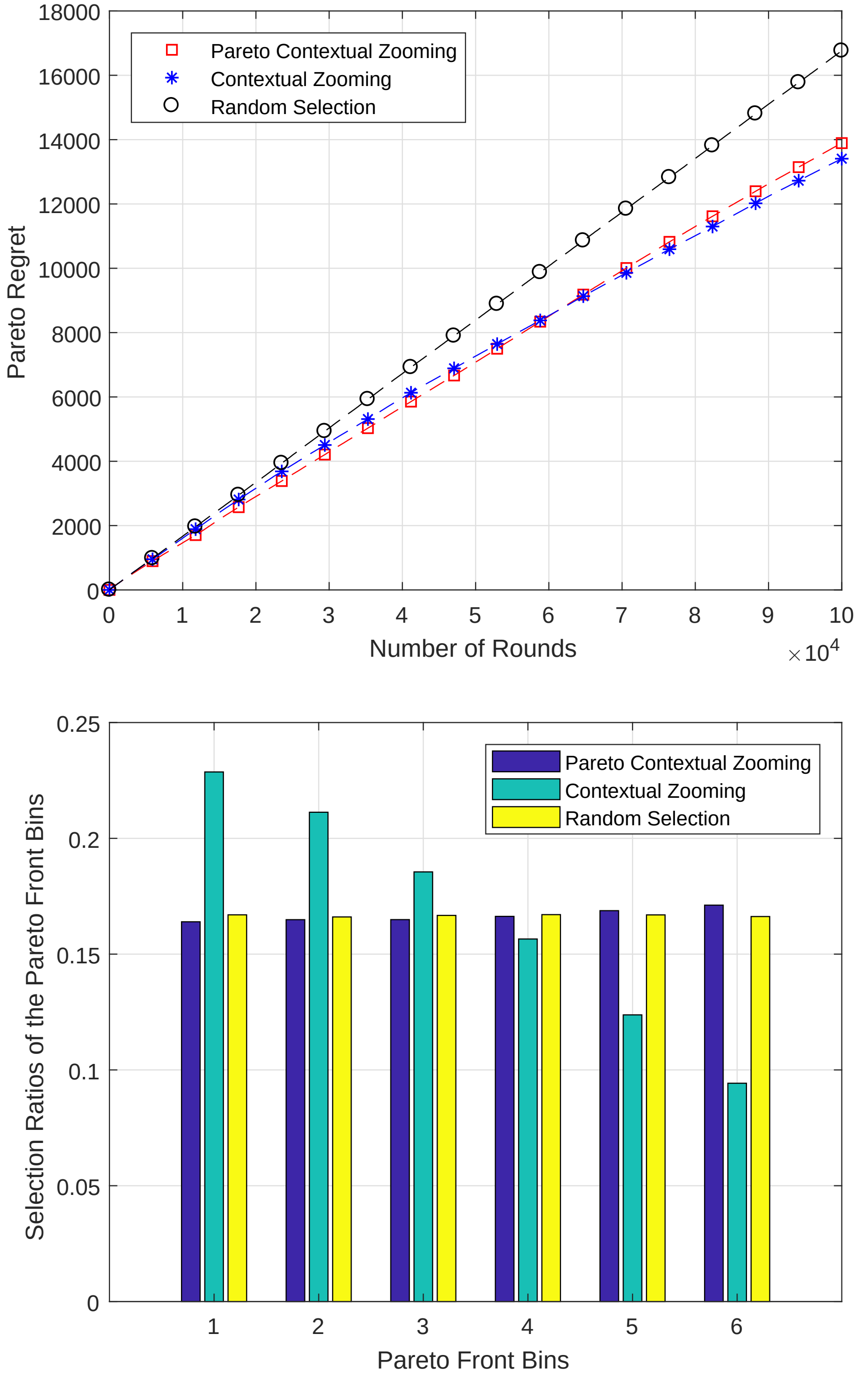

Figure 4.2: (i) Pareto Regret vs. Number of Rounds (ii) Selection Ratio of the Pareto Front Bins

Chapter 5

Conclusion and Future Works

In this thesis, we consider two different multi-objective contextual MAB problems. In the first one, we propose a novel contextual MAB problem with two objectives in which one objective is dominant and the other is non-dominant. We also propose a new performance metric called 2D regret which is a vector whose ith component corresponds to the difference between the expected total reward of an oracle in objective i that selects optimal arm for each context and that of the learner by round T. We extend the Pareto regret proposed in [10] to take into account the dependence of the Pareto front on the context. For this problem, we propose a multi-objective contextual multi-armed bandit algorithm (MOC-MAB) and we prove that both the 2D regret and the Pareto regret of MOC-MAB are sublinear in the number of rounds. However, MOC-MAB partitions the context space uniformly and it is not an optimal solution when context sampling distribution is complex. For the future, we will work on this problem by considering adaptive partition and by this way, our initial assumption is to provide tighter regret bounds in the 2D and the Pareto regret. In the second work of the thesis, we consider a multi-objective contextual MAB problem with an arbitrary number of objectives and a high-dimensional, possibly uncountable arm set. Then, we propose an online learning algorithm called Pareto Contextual Zooming (PCZ) and prove that Pareto regret of PCZ is sublinear in the number of rounds. We also show an almost matching lower bound $\Omega(T^{(1+d_p)/(2+d_p)})$ based on a reduction to the classical contextual bandit problem with similarity information, which shows that our bound is tight up to logarithmic factors. Performance of PCZ is demonstrated in the simulations and we show that PCZ randomly alternates between the arms in the estimated Pareto front and ensures the arms in this set are fairly selected. Future work will focus on evaluating the Pareto regret and the fairness of the proposed algorithm in real-world datasets.

Bibliography

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” in *Proc. 25th Conf. in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[2] D. Russo and B. Van Roy, “Learning to optimize via posterior sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[3] H. Robbins, “Some aspects of the sequential design of experiments,” in *Herbert Robbins Sel. Papers*, pp. 169–177, Springer, 1985.

[4] Y. Gai, B. Krishnamachari, and R. Jain, “Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation,” in *Proc. 4th IEEE Symposium on New Frontiers in Dynamic Spectrum*, pp. 1–9, 2010.

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

[6] C. Tekin, J. Yoon, and M. van der Schaar, “Adaptive ensemble learning with confidence bounds,” *IEEE Transactions on Signal Processing*, vol. 65, no. 4, pp. 888–903, 2017.

[7] A. Slivkins, “Contextual bandits with similarity information,” *J. of Machine Learning Research*, vol. 15, no. 1, pp. 2533–2568, 2014.

[8] T. Lu, D. Pál, and M. Pál, “Contextual multi-armed bandits,” in *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics*, pp. 485–492, 2010.

[9] M. Rodriguez, C. Posse, and E. Zhang, "Multiple objective optimization in recommender systems," in *Proc. 6th ACM Conf. on Recommender Systems*, pp. 11–18, 2012.

[10] M. M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *Proc. 23rd Int. Joint Conf. on Neural Networks*, pp. 1–8, 2013.

[11] C. Tekin and E. Turgay, "Multi-objective contextual multi-armed bandit with a dominant objective," *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3799–3813, 2018.

[12] E. Turgay, D. Oner, and C. Tekin, "Multi-objective contextual bandit problem with similarity information," in *Proc. 21st. Int. Conf. on Artificial Intelligence and Statistics*, pp. 1673–1681, 2018.

[13] C. Tekin and M. van der Schaar, "Distributed online learning via cooperative contextual bandits," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3700–3714, 2015.

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

[15] W. Jouini, D. Ernst, C. Moy, and J. Palicot, "Multi-armed bandit based policies for cognitive radio's decision making issues," in *Proc. 3rd Int. Conf. on Signals, Circuits and Systems*, 2009.

[16] K. Liu and Q. Zhao, "Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players," in *Proc. 35th IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 3010–3013, IEEE, 2010.

[17] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 300–309, 2012.

[18] M. A. Qureshi and C. Tekin, “Online cross-layer learning in heterogeneous cognitive radio networks without csi,” in *Proc. 26th Signal Processing and Communications Applications Conf.*, pp. 1–4, IEEE, 2018.

[19] C. Tekin and E. Turgay, “Multi-objective contextual bandits with a dominant objective,” in *Proc. 27th IEEE Int. Workshop on Machine Learning for Signal Processing*, pp. 1–6, IEEE, 2017.

[20] W. Chu, Li, L. Reyzin, and R. E. Schapire, “Contextual bandits with linear payoff functions,” in *Proc. 14th Int. Conf. on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

[21] J. Langford and T. Zhang, “The epoch-greedy algorithm for contextual multi-armed bandits,” *Proc. Advances in Neural Information Processing Systems*, vol. 20, pp. 1096–1103, 2007.

[22] 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. on Machine Learning*, pp. 1638–1646, 2014.

[23] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[24] R. Agrawal, “Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem,” *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[25] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[26] A. Garivier and O. Cappé, “The KL-UCB algorithm for bounded stochastic bandits and beyond,” in *Proc. 24th Annual Conf. on Learning Theory*, pp. 359–376, 2011.

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

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

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

[30] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang, "Efficient optimal learning for contextual bandits," in *Proc. 27th Conf. on Uncertainty in Artificial Intelligence*, 2011.

[31] S. Q. Yahyaa, M. M. Drugan, and B. Manderick, "Knowledge gradient for multi-objective multi-armed bandit algorithms.," in *Proc. 6th Int. Conf. on Agents and Artificial Intelligence*, pp. 74–83, 2014.

[32] S. Q. Yahyaa and B. Manderick, "Thompson sampling for multi-objective multi-armed bandits problem," in *Proc. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 47–52, 2015.

[33] S. Q. Yahyaa, M. M. Drugan, and B. Manderick, "Annealing-Pareto multi-objective multi-armed bandit algorithm," in *Proc. 4th IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 1–8, 2014.

[34] M. M. Drugan and A. Nowé, "Scalarization based Pareto optimal set of arms identification algorithms," in *Proc.24th Int. Joint Conf. on Neural Networks*, pp. 2690–2697, 2014.

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

[36] S. Mannor and N. Shimkin, "A geometric approach to multi-criterion reinforcement learning," *J. of Machine Learning Research*, vol. 5, no. Apr, pp. 325–360, 2004.

[37] M. Ehrgott, *Multicriteria optimization*, vol. 491. Springer Science & Business Media, 2005.

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

[39] D. T. Hoang, E. L. Linzer, and J. S. Vitter, "Lexicographic bit allocation for mpeg video," *J. of Visual Communication and Image Representation*, vol. 8, no. 4, pp. 384–404, 1997.

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

[41] 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. 21st Association for the Advancement of Artificial Intelligence*, vol. 6, pp. 1630–1633, 2006.

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

[43] V. Shah-Mansouri, A.-H. Mohsenian-Rad, and V. W. Wong, "Lexicographically optimal routing for wireless sensor networks with multiple sinks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1490–1500, 2009.

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

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

[46] W. Chu, S.-T. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah, "A case study of behavior-driven conjoint analysis on yahoo!:

front page today module," in *Proc. 15th ACM Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1097–1104, ACM, 2009.

Appendix A

Concentration Inequality [1, 2] - MOC-MAB

Consider an arm a for which the rewards of objective i are generated by a process $\{R^i_a(t)\}_{t=1}^{T}$ with $\mu^i_a = \mathrm{E}[R^i_a(t)]$, 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 the 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

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

Appendix B

Concentration Inequality [1, 2] - PCZ

Consider a ball B for which the rewards of objective i are generated by a process $\{R^i_B(t)\}_{t=1}^T$ with mean $\mu^i_B = \mathrm{E}[R^i_B(t)]$, where the noise $R^i_B(t) - \mu^i_B$ is 1-sub-Gaussian. Recall that $B(t)$ is the ball selected in round t and $N_B(T)$ is the number of times ball B is selected by the beginning of round T. Let $\hat{\mu}^i_B(T) = \sum_{t=1}^{T-1} \mathrm{I}(B(t) = B)R^i_B(t)/N_B(T)$ for $N_B(T) > 0$ and $\hat{\mu}^i_B(T) = 0$ for $N_B(T) = 0$. Then, for any $0 < \theta < 1$ with probability at least $1 - \theta$ we have

$$\left|\hat{\mu}^i_B(T) - \mu^i_B\right| \leq \sqrt{\frac{2}{N_B(T)}\left(1 + 2\log\left(\frac{(1 + N_B(T))^{1/2}}{\theta}\right)\right)} \quad \forall T \in \mathbb{N}.$$

Appendix C

Table of Notation

Table C.1: Table of Notation

Notation	Description
x_t	d_x-dimensional context vector observed by the learner at the beginning of round t
$\mathcal{X}$	Context set
d_x	Number of dimensions of $\mathcal{X}$
a_t	Arm chosen by the learner at round t
$\mathcal{A}$	Arm set
d_a	Number of dimensions of $\mathcal{A}$
r_t^i	Reward obtained from objective $i \in \{1, \dots, d_r\}$ in round t.
d_r	Number of objectives
$\boldsymbol{r}_t$ (or r_t)	d_r-dimensional reward vector obtained in round t, i.e. $\boldsymbol{r}_t := (r_t^i, \dots, r_t^{d_r})$
$\mu_a^i(x)$	Expected reward of arm $a \in \mathcal{A}$ in objective $i \in \{1, \dots, d_r\}$, given context $x \in \mathcal{X}$.
$\boldsymbol{\mu}_a(x)$ (or $\mu_a(x)$)	d_r-dimensional expected reward vector of arm $a \in \mathcal{A}$, given context $x \in \mathcal{X}$. , i.e. $\boldsymbol{\mu}_a(x) := (\mu_a^i(x), \dots, \mu_a^{d_r}(x))$
κ_t^i	Noise process obtained from objective $i \in \{1, \dots, d_r\}$ in round t.
$\boldsymbol{\kappa}_t$ (or κ_t)	d_r-dimensional noise process vector obtained in round t, i.e. $\boldsymbol{\kappa}_t := (\kappa_t^i, \dots, \kappa_t^{d_r})$
$\mathcal{A}^*(x)$	Set of arms that maximize the expected reward for the dominant objective for context x
$a^*(x)$	Optimal arm (for multi-objective contextual Bandit with a dominant objective problem formulation) given a context $x \in \mathcal{X}$, which 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
$\mathcal{O}(x)$	Set of all Pareto optimal arms (also called the Pareto front), for given a particular context x,
$\text{Reg}^1(t)$	Cumulative 2D regret in the dominant objective until round t
$\text{Reg}^2(t)$	Cumulative 2D regret in the non-dominant objective until round t
$\text{PR}(t)$	Cumulative Pareto regret until round t
$\Delta_a(x)$,	Pareto suboptimality gap of an arm $a \in \mathcal{A}$ given context $x \in \mathcal{X}$.
p^*	Partition such that $x_t \in p^*$.
$\mathcal{P}$	Partition set
$\hat{\mu}^1_{a,p}$	Sample mean of the rewards obtained from rounds prior to round t in which the context was in $p \in \mathcal{P}$ and arm $a \in \mathcal{A}$ was selected for the dominant objective.
$\hat{\mu}^2_{a,p}$	Sample mean of the rewards obtained from rounds prior to round t in which the context was in $p \in \mathcal{P}$ and arm $a \in \mathcal{A}$ was selected for the non-dominant objective.
$\hat{A}^*$	Candidate optimal arm set chosen by MOC-MAB

Notation	Description
$N_{a,p}(t)$	Number of times the context was in $p \in \mathcal{P}$ and arm $a \in \mathcal{A}$ was selected before round t.
$g^1_{a,p}$	Index of arm $a \in \mathcal{A}$ for the rewards in the dominant objective for partition $p \in \mathcal{P}$
$g^2_{a,p}$	Index of arm $a \in \mathcal{A}$ for the rewards in the non-dominant objective for partition $p \in \mathcal{P}$
$B(t)$	Ball selected by PCZ in round t
$\mathcal{B}(t)$	Set of balls created by PCZ by the beginning of round t.
$\hat{R}(x_t)$	Set of relevant balls, i.e. $\hat{\mathcal{R}}(x_t) := \{B \in \mathcal{B} : (x_t, a) \in \mathrm{dom}_t(B)$ for some $a \in \mathcal{A}\}$.
$\hat{B}^*$	Candidate Pareto front set among the set of balls in $\hat{R}(x_t)$ chosen by PCZ
$g_B^{i,pre}(t)$	Pre-index of ball $B \in \mathcal{B}$ in objective $i \in 1, \ldots, d_r$ at round t
$g_B^i(t)$	Index of ball $B \in \mathcal{B}$ in objective $i \in 1, \ldots, d_r$ at round t
$\boldsymbol{g}_B(t)$	d_r-dimensional index vector of ball $B \in \mathcal{B}$ at round t, i.e. $\boldsymbol{g}_B(t) \coloneqq (g_B^1(t), \ldots, g_B^{d_r}(t))$
$N_B(t)$	Number of times ball $B \in \mathcal{B}$ was selected before round t.
$\hat{\mu}_B^i(t)$	Sample mean of the rewards for objective $i \in \{1, \ldots\}$, obtained from rounds prior to round t in which ball $B \in \mathcal{B}$ was selected
x_B	Center of ball $B \in \mathcal{B}$ for the context dimensions
a_B	Center of ball $B \in \mathcal{B}$ for the arm dimensions
$r(B)$	Radius of ball $B \in \mathcal{B}$
d_p	Number of Pareto zooming dimension