Multi-objective Contextual Bandit Problem with Similarity Information

Eralp Turğay
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

Doruk Öner
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

Cem Tekin
Electrical and Electronics Engineering Department
Bilkent University
Ankara, Turkey

Abstract

In this paper we propose the multi-objective contextual bandit problem with similarity information. This problem extends the classical contextual bandit problem with similarity information by introducing multiple and possibly conflicting objectives. Since the best arm in each objective can be different given the context, learning the best arm based on a single objective can jeopardize the rewards obtained from the other objectives. In order to evaluate the performance of the learner in this setup, we use a performance metric called the contextual Pareto regret. Essentially, the contextual Pareto regret is the sum of the distances of the arms chosen by the learner to the context dependent Pareto front. For this problem, we develop a new online learning algorithm called Pareto Contextual Zooming (PCZ), which exploits the idea of contextual zooming to learn the arms that are close to the Pareto front for each observed context by adaptively partitioning the joint context-arm set according to the observed rewards and locations of the context-arm pairs selected in the past. Then, we prove that PCZ achieves $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ Pareto regret where d_p is the Pareto zooming dimension that depends on the size of the set of near-optimal context-arm pairs. Moreover, we show that this regret bound is nearly optimal by providing an almost matching $\Omega(T^{(1+d_p)/(2+d_p)})$ lower bound.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

1 INTRODUCTION

The multi-armed bandit (MAB) is extensively used to model sequential decision making problems with uncertain rewards. While many real-world applications ranging from cognitive radio networks [1] to recommender systems [2] to medical diagnosis [3] require intelligent decision making mechanisms that learn from the past, majority of these applications involve side-observations that can guide the decision making process, which does not fit into the classical MAB model. This issue is resolved by proposing new MAB models, called contextual bandits, that learn how to act optimally based on side-observations [2, 4, 5]. 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 [6], and sensitivity and specificity in medical diagnosis.

This motivates us to develop a new MAB model that addresses the learning challenges that arise from side-observations and presence of multiple objectives at the same time. We call this model the multi-objective contextual bandit problem with similarity information. In this problem, at the beginning of each round, the learner observes a context from the context set $\mathcal{X}$ and selects an arm from the arm set $\mathcal{Y}$. At the end of the round, the learner observes a random reward whose distribution depends on the observed context and the selected arm. We assume that the sequence of contexts is fixed beforehand and does not depend on the decision of the learner. Other than this, we impose no assumptions on how the contexts are sampled from $\mathcal{X}$, but we assume that the expected reward of arm y given context x is time-invariant. The goal of the learner is to maximize its total reward in each objective (while ensuring some sort of fairness, which will be described later) over T rounds by judiciously selecting good arms only based on the past observations and selections. To

facilitate learning, we also assume that the learner is endowed with similarity information, which 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 merely states that the expected reward function is Lipschitz continuous.

Different from the classical contextual bandit problem with similarity information [4] in which the reward is scalar, in our problem the reward is multi-dimensional. In order to measure how well the learner performs with respect to the oracle that perfectly knows the expected reward of each context-arm pair, we adopt the notion of contextual Pareto regret defined in [7] for two objectives, and extend it to work for an arbitrary number of objectives. The contextual Pareto regret (referred to as the Pareto regret hereafter) measures the sum of the distances of the expected rewards of the arms chosen by the learner to the Pareto front given the contexts. Importantly, the 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. Moreover, 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 by proposing an online learning algorithm called Pareto Contextual Zooming (PCZ) that is built on the contextual zooming algorithm in [4], and show that it achieves sublinear Pareto regret. Essentially, we provide a finite-time bound for the Pareto regret with a time order $\tilde{O}(T^{(1+d_p)/(2+d_p)})$, 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. The proposed algorithm achieves this regret bound by adaptively partitioning the similarity space according to the empirical distribution of the past arm selections, and context and reward observations. While doing so, it only needs to keep track of a set of parameters defined for an active set of balls that cover the similarity space, and it only needs to find the balls which are not Pareto dominated by any other ball among all balls that are relevant to the current round. This shows that by making use of the similarity information, a complete characterization of the Pareto front is not necessary to achieve sublinear regret. Finally, we 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.

Rest of the paper is organized as follows. Related work is described in Section 2. Problem formulation is given in Section 3. The learning algorithm is proposed in Section 4, and its Pareto regret is upper bounded in Section 5. Section 6 provides a lower bound on the Pareto regret. Section 7 concludes the paper. Due to limited space, some of the technical proofs and numerical results are given in the supplemental document.

2 RELATED WORK

We split the discussion on related work into three parts: related work in contextual bandits, related work in multi-objective bandits and related work that considers multi-objective contextual bandits.

Many different formulations of the contextual bandit problem exist in the literature, which leads to different types of algorithms and regret bounds. For instance, [10] considers the problem where the contexts and rewards are sampled from a time-invariant distribution, and proposes the epoch greedy algorithm which achieves $O(T^{2/3})$ regret. Later on, more efficient algorithms [11, 12] that achieve $\tilde{O}(T^{1/2})$ regret are developed for this problem.

Another line of research focuses on contextual bandit problems under the linear realizability assumption, which requires the expected reward of an arm to be linear in its features (or equivalently contexts). Numerous learning algorithms are proposed for this problem including LinUCB [2] and SuplinUCB [8] that achieve $\tilde{O}(\sqrt{Td})$ regret, where d is the context dimension. SuplinUCB is also extended to work with kernel functions in [13], and is shown to achieve $\tilde{O}(\sqrt{T\tilde{d}})$ regret, where $\tilde{d}$ represents the effective dimension of the kernel feature space. However, these works do not take collaborative information into account. Some of the contexts may share similar behavior given the same arm and constitute clusters. In this case, the learner does not need to learn a different model for each context, but just a single model for each cluster. In [14], CLUB uses confidence balls of the clusters (contexts are related to users in their work) to both estimate user similarity, and to learn jointly for users within the same cluster. An extension of this work [15] considers the "two-sided clustering" where clustering process is done by simultaneously grouping arms based on similarity at the context side and contexts based on similarity at the arm side.

Our work builds on the formalism of contextual bandits with similarity information. The prior work in this category attempts to minimize the regret without imposing any stochastic assumptions on the context arrivals. However, knowledge of the similarity

Table 1: Comparison with Related Work

Bandit algorithm	Regret bound	Multi-objective	Contextual	Linear rewards	Similarity assumption	Adaptive partition
Contextual Zooming [4]	$\tilde{O}(T^{1-1/(2+d_z)})$	No	Yes	No	Yes	Yes
Query-Ad-Clustering [5]	$\tilde{O}(T^{1-1/(2+d_c)})$	No	Yes	No	Yes	No
SupLinUCB [8]	$\tilde{O}(\sqrt{T})$	No	Yes	Yes	No	No
Pareto-UCB1 [9]	$O(\log(T))$ (Pareto regret)	Yes	No	No	No	No
Scalarized-UCB1[9]	$O(\log(T))$	Yes	No	No	No	No
MOC-MAB [7]	$\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ (Pareto regret)	Yes	Yes	No	Yes	No
PCZ (this paper)	$\tilde{O}(T^{1-1/(2+d_p)})$ (Pareto regret)	Yes	Yes	No	Yes	Yes

information, which relates the distances between expected rewards to the distances between the context-arm pairs in the similarity space, is required. With this assumption, Query-Ad-Clustering algorithm in [5] achieves $O(T^{1-1/(2+d_c)+\epsilon})$ regret for any $\epsilon > 0$, where d_c is the covering dimension of the similarity space. This algorithm works by partitioning the similarity space into disjoint sets and estimating the expected arm rewards for each set of the partition separately. Another related work proposes the contextual zooming algorithm [4] that judiciously partitions the similarity space based on the reward structure of the similarity space and past context arrivals. It is shown that adaptive partitioning achieves better performance compared to uniform partitioning since it uses the experience gained in the past to form more accurate reward estimates in the potentially rewarding regions of the similarity space. While PCZ borrows the idea of contextual zooming from [4], the arm selection strategy, the regret notion, and the analysis of the regret of PCZ are substantially different from the prior work. Notably, the regret bound of the contextual zooming algorithm in [4] depends on the zooming dimension d_z, while our regret bound depends on the Pareto zooming dimension d_p. In our case, d_z depends on the objective selected by the contextual zooming algorithm. While, in general, d_z is smaller than d_p,[1] it is impossible to guarantee fairness over context-arm pairs in the Pareto front by using the contextual zooming algorithm. In addition, due to the multi-dimensional nature of the problem, different bounding techniques are required to show that PCZ selects arms that are near the Pareto front with high probability. Moreover, we make use of a different concentration inequality that allows us to deal with noise processes that are conditionally 1-sub-Gaussian, and also provide finite-time bounds on the Pareto regret and the expected Pareto regret.

Learning with multiple objectives is mainly investigated under the classical MAB model without contexts. Numerous algorithms are proposed to minimize the Pareto regret, which measures the total loss due to selecting arms that are not in the Pareto front. For this problem, an algorithm called ParetoUCB1 is shown to achieve $O(\log T)$ Pareto regret [9]. This algorithm simply uses the UCB indices instead of the expected arm rewards to select an arm from the estimated Pareto front in each round. Other algorithms include the Pareto Thompson sampling [16], the Annealing Pareto [16] and the Pareto-KG [17]. Another line of research aims to scalarize the multi-objective problem by assigning weights to each objective. For instance, Scalarized UCB1 [9] learns the best scalarization functions among a given set of scalarization functions, and achieves $O(S' \log(T/S'))$ scalarized regret where S' is the number of scalarization functions used by the algorithm. In another work [18], the hierarchical optimistic optimization strategy for the $\mathcal{X}$-armed bandit problem [19] is extended to the multi-objective setup, but theoretical regret analysis of the proposed method is not given.

Apart from the above works, a specific multi-objective contextual bandit problem with dominant and non-dominant objectives and finite number of arms is considered in [7]. They define the regret in each objective separately with respect to a benchmark that always picks a specific arm in the Pareto front that favors the reward in the dominant objective over the reward in the non-dominant objective. For this problem, they propose a contextual bandit algorithm that uniformly partitions the context set, and prove that it achieves $\tilde{O}(T^{(\alpha+d)/(2\alpha+d)})$ Pareto regret, where d is the dimension of the context and α is a similarity information dependent constant. Our work significantly differs from [7] since we consider a very general similarity space and an adaptive contextual zooming algorithm. Moreover, we do not prioritize the objectives. A detailed comparison of our work with the related works is given in Table 1.

[1] Implicitly, both d_z and d_p are functions of a constant $\tilde{c}$ that is related to an r-packing of a subset of the similarity space, and the regret bounds hold for any value of $\tilde{c} > 0$. Here, we compare d_z and d_p given the same constant.

3 PROBLEM DESCRIPTION

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 y_t from a d_y-dimensional arm set $\mathcal{Y}$. 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 $\mu_y^i(x)$ denote the expected reward of arm y in objective i for context x and $\mu_y(x) := (\mu_y^1(x), \ldots, \mu_y^{d_r}(x))$. The random reward vector obtained from arm y_t in round t is given as $r_t := \mu_{y_t}(x_t) + \kappa_t$ where κ_t is the d_r-dimensional noise process whose marginal distribution for each objective is conditionally 1-sub-Gaussian, i.e., $\forall \lambda \in \mathbb{R}$

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

where $b_{1:t} := (b_1, \ldots, b_t)$. Context and arm sets together constitute the set of feasible context-arm pairs, denoted by $\mathcal{P} := \mathcal{X} \times \mathcal{Y}$. 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 1. *For all $i \in \{1, \ldots, d_r\}$, $y, y' \in \mathcal{Y}$ and $x, x' \in \mathcal{X}$, we have*

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

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

$(\mathcal{P}, D)$ 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 1 (Pareto optimality)**.** *(i) An arm y is* weakly dominated *by arm y' given context x, denoted by $\mu_y(x) \preceq \mu_{y'}(x)$ or $\mu_{y'}(x) \succeq \mu_y(x)$, if $\mu_y^i(x) \leq \mu_{y'}^i(x), \forall i \in \{1, \ldots, d_r\}$.*
(ii) An arm y is dominated *by arm y' given context x, denoted by $\mu_y(x) \prec \mu_{y'}(x)$ or $\mu_{y'}(x) \succ \mu_y(x)$, if it is weakly dominated and $\exists i \in \{1, \ldots, d_r\}$ such that $\mu_y^i(x) < \mu_{y'}^i(x)$.*
(iii) Two arms y and y' are incomparable given context x, denoted by $\mu_y(x) || \mu_{y'}(x)$, 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 2 (PSG)**.** *The PSG of an arm $y \in \mathcal{Y}$ given context x, denoted by $\Delta_y(x)$, is defined as the minimum scalar $\epsilon \geq 0$ that needs to be added to all entries of $\mu_y(x)$ such that y becomes a member of the Pareto front. Formally,*

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

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

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} \Delta_{y_t}(x_t). \tag{1}$$

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 3. *(i) $\hat{\mathcal{P}} \subset \mathcal{P}$ is called an r-packing of $\mathcal{P}$ if all $z, z' \in \hat{\mathcal{P}}$ satisfies $D(z, z') \geq r$. For any $r > 0$, the r-packing number of $\mathcal{P}$ is $A_r^{packing}(\mathcal{P}) := \max\{|\hat{\mathcal{P}}| : \hat{\mathcal{P}}$ is an r-packing of $\mathcal{P}\}$.*
(ii) For a given $r > 0$, let $\mathcal{P}_{\mu,r} := \{(x, y) \in \mathcal{P} : \Delta_y(x) \leq 12r\}$ denote the set of near-optimal context-arm pairs. The Pareto r-zooming number N_r is defined as the r-packing number of $\mathcal{P}_{\mu,r}$.
(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 THE LEARNING ALGORITHM

In this section, we present a multi-objective contextual bandit algorithm called *Pareto Contextual Zooming* (PCZ). Pseudo-code of PCZ is given in Algorithm 1. PCZ is a multi-objective extension of the single objective contextual zooming algorithm [4]. 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, y_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

Algorithm 1 Pareto Contextual Zooming

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

B that excludes all active balls at the beginning of round t that have radius strictly smaller than $r(B)$, i.e., $\text{dom}_t(B) := B \setminus (\cup_{B' \in \mathcal{B}: r(B') < r(B)} B')$. The domains of all active balls cover the similarity space.

Initially, PCZ takes as inputs the time horizon T,[2] the similarity space $(\mathcal{P}, D)$, 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 $\hat{\mathcal{R}}(x_t) := \{B \in \mathcal{B} : (x_t, y) \in \text{dom}_t(B) \text{ for some } y \in \mathcal{Y}\}$.

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

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

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

which is the sum of the sample mean reward $\hat{\mu}_B^i$, 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

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

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 1), which is given as $\hat{\mathcal{A}}^* := \{B \in \hat{\mathcal{R}}(x_t) : g_B \not\prec g_{B'}, \forall B' \in \hat{\mathcal{R}}(x_t)\}$, 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, y_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 over-

[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.

come 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, y_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_B^i is used instead of $g_B^{i,pre}$ as an enhanced UCB, which is in general tighter than the UCB based on the pre-index.

5 REGRET ANALYSIS

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 $\mathcal{F}$, let $\mathcal{F}^c$ 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 $R_B^i(t)$ 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, $\text{dom}_t(B) \subseteq \text{dom}(B)$, $\forall t \in \{\tau_B(0), \dots, T\}$. For $B \in \mathcal{B}'(T)$, let $\mathcal{T}_B$ denote the set of rounds in $\{\tau_B(0), \dots, T\}$ in which ball B is selected by PCZ. Also let $\tilde{x}_B(t) := x_{\tau_B(t)-1}$, $\tilde{y}_B(t) := y_{\tau_B(t)-1}$, $\tilde{R}_B^i(t) := R_B^i(\tau_B(t)-1)$, $\tilde{\kappa}_B^i(t) := \kappa_{\tau_B(t)-1}^i$, $\tilde{N}_B(t) := N_B(\tau_B(t))$, $\tilde{\mu}_B^i(t) := \hat{\mu}_B^i(\tau_B(t))$, $\tilde{g}_B^{i,pre}(t) := g_B^{i,pre}(\tau_B(t))$, $\tilde{g}_B^i(t) := g_B^i(\tau_B(t))$, and $\tilde{u}_B(t) := u_B(\tau_B(t))$. We note that all inequalities that involve random variables hold with probability one unless otherwise stated.

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

Let

$$\text{Reg}_B(T) := \sum_{t=1}^{N_B(T+1)} \Delta_{\tilde{y}_B(t)}(\tilde{x}_B(t))$$

denote the Pareto regret incurred in ball $B \in \mathcal{B}'(T)$ for rounds in $\mathcal{T}_B$. Then, the Pareto regret in (1) can be written as

$$\text{Reg}(T) = \sum_{B \in \mathcal{B}'(T)} \text{Reg}_B(T).$$

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

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

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

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

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

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

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

$$\begin{aligned}\text{E}[\text{Reg}(T)] &= \text{E}[\text{Reg}(T) \mid \text{UC}] \Pr(\text{UC}) \\ &\quad + \text{E}[\text{Reg}(T) \mid \text{UC}^c] \Pr(\text{UC}^c) \\ &\le C_{\max} T \Pr(\text{UC}) + \text{E}[\text{Reg}(T) \mid \text{UC}^c] \quad (2)\end{aligned}$$

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

$$\Pr(\text{UC}) = \int \Pr(\text{UC} | \mathcal{B}(T)) dQ(\mathcal{B}(T)) \quad (3)$$

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

$$\begin{aligned}\Pr(\text{UC}|\mathcal{B}(T)) &= \Pr(\cup_{B\in\mathcal{B}(T)}\text{UC}_B|\mathcal{B}(T)) \\ &\leq \sum_{B\in\mathcal{B}(T)}\Pr(\text{UC}_B|\mathcal{B}(T)). \end{aligned} \tag{4}$$

We proceed by bounding the term Pr(UC) in (2). For this, first we bound $\Pr(\text{UC}_B|\mathcal{B}(T))$ in the next lemma, whose proof is given in the supplemental document.

Lemma 1. *When PCZ is run, we have* $\Pr(UC_B|\mathcal{B}(T)) \leq \delta/T$, $\forall B \in \mathcal{B}(T)$.

Next, we bound Pr(UC). Since $|\mathcal{B}(T)| \leq T$, by using union bound over all created balls, (3) and (4) we obtain

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

Then, by using (2),

$$\begin{aligned}\text{E}[\text{Reg}(T)] &\leq C_{\max}T\Pr(\text{UC}) + \text{E}[\text{Reg}(T)\mid \text{UC}^c] \\ &\leq C_{\max}T\delta + \text{E}[\text{Reg}(T)\mid \text{UC}^c]. \end{aligned} \tag{5}$$

In the remainder of the analysis we bound Reg(T) on event UC^c. Then, we conclude the analysis by using this to bound (5). For simplicity of notation, in the following lemmas we use $B(t)$ to denote the ball selected in round t.

Lemma 2. *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*

$$g_{B(t)}(t) \not\prec \mu_y(x_t), \forall y \in \mathcal{Y}.$$

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

$$\begin{aligned}g^i_B(t) &= r(B) + g^{i,pre}_{\tilde{B}_i}(t) + D(B,\tilde{B}_i) \\ &= r(B) + \hat{\mu}^i_{\tilde{B}_i}(t) + u_{\tilde{B}_i}(t) + r(\tilde{B}_i) + D(B,\tilde{B}_i) \\ &= r(B) + U^i_{\tilde{B}_i}(t) + r(\tilde{B}_i) + D(B,\tilde{B}_i) \\ &\geq r(B) + \mu^i_{y_{\tilde{B}_i}}(x_{\tilde{B}_i}) + D(B,\tilde{B}_i) \\ &\geq r(B) + \mu^i_{y_B}(x_B) \geq \mu^i_y(x_t)\end{aligned}$$

where the first inequality holds by the definition of UC^c, and the second and the third inequalities hold by Assumption 1. According to the above inequality $g_B(t) \succeq \mu_y(x_t)$, $\forall y$ such that $(x_t,y) \in \text{dom}_t(B)$. We also know that $\forall y \in \mathcal{Y}$, $\exists B \in \hat{\mathcal{R}}(x_t)$ such that $(x_t,y) \in \text{dom}_t(B)$. 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, $g_{B(t)}(t) \not\prec \mu_y(x_t), \forall y \in \mathcal{Y}$, and hence, the virtual arm with expected reward vector $g_{B(t)}(t)$ is not dominated by any of the arms. □

Lemma 3. *When PCZ is run, on event UC^c, we have*

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

Proof. This proof is similar to the proof in [4]. 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

$$\begin{aligned}g^{i,pre}_{B^{par}(t)}(t) &= \hat{\mu}^i_{B^{par}(t)}(t) + r(B^{par}(t)) + u_{B^{par}(t)}(t) \\ &= L^i_{B^{par}(t)}(t) + r(B^{par}(t)) + 2u_{B^{par}(t)}(t) \\ &\leq \mu^i_{y_{B^{par}(t)}}(x_{B^{par}(t)}) \\ &\quad + 2r(B^{par}(t)) + 2u_{B^{par}(t)}(t) \\ &\leq \mu^i_{y_{B^{par}(t)}}(x_{B^{par}(t)}) + 4r(B^{par}(t)) \\ &\leq \mu^i_{y_{B(t)}}(x_{B(t)}) + 5r(B^{par}(t)) \end{aligned} \tag{6}$$

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 1. We also have

$$\begin{aligned}g^i_{B(t)}(t) &\leq r(B(t)) + g^{i,pre}_{B^{par}(t)}(t) + D(B^{par}(t),B(t)) \\ &\leq r(B(t)) + g^{i,pre}_{B^{par}(t)}(t) + r(B^{par}(t)) \\ &\leq r(B(t)) + \mu^i_{y_{B(t)}}(x_{B(t)}) + 6r(B^{par}(t)) \\ &\leq \mu^i_{y_{B(t)}}(x_{B(t)}) + 13r(B(t)) \\ &\leq \mu^i_{y_t}(x_t) + 14r(B(t))\end{aligned}$$

where the third inequality follows from (6). Since $g^i_{B(t)}(t) - \mu^i_{y_t}(x_t) \leq 14r(B(t))$ for all $i \in \{1,\dots,d_r\}$ and the virtual arm is not dominated by any arm in the Pareto front by Lemma 2, the PSG of the selected arm is bounded by $\Delta_{y_t}(x_t) \leq 14r(B(t))$. □

Lemma 4. *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 3. Moreover, in any round t in which a radius r ball is created, we have $\Delta_{y_t}(x_t) \leq 12r$.*

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

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,

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

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 1, and the last inequality follows from $r(B'(t)) = r(B(t))/2$. Similar to the proof of Lemma 3, since $g^i_{B(t)}(t) - \mu^i_{y_{B'(t)}}(x_{B'(t)}) \leq 12r(B'(t))$ for all $i \in \{1, \dots, d_r\}$ and the virtual arm is not dominated by any arm in the Pareto front by Lemma 2, the PSG of point $(x_{B'(t)}, y_{B'(t)}) = (x_t, y_t)$ 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 $\mathcal{P}_{\mu,r(B'(t))}$. 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 $\mathcal{P}_{\mu,r}$, which is N_r. □

The following lemma (proof is given in the supplemental document), bounds the regret of PCZ in terms of N_r by using the results in Lemmas 3 and 4.

Lemma 5. *On event UC^c, the Pareto regret of PCZ by round T is bounded by $Reg(T) \leq 28Tr_0 + \sum_{r=2^{-i}:i\in\mathbb{N},r_0\leq r\leq 1} 56r^{-1}N_r \log(2\sqrt{2}d_r T^{\frac{3}{2}}e/\delta)$ for any $r_0 \in (0,1]$.*

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

Theorem 1. *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*

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

where d_p is given in Definition 3.

Proof. Using Definition 3 and the result of Lemma 5 on event UC^c, we have

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

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

Corollary 1. *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*

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

Proof. Theorem 1 is used in (5) with $\delta = 1/T$. □

6 LOWER BOUND

It is shown in [4] 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 $\mu^i_y(x) = \mu^j_y(x)$, $\forall (x,y) \in \mathcal{P}$ and $\forall i,j \in \{1,\dots,d_r\}$, and $\kappa^i_t = \kappa^j_t$, $\forall t \in \{1,\dots,T\}$ and $\forall i,j \in \{1,\dots,d_r\}$. 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.

7 CONCLUSION

In this paper we propose the multi-objective contextual bandit problem which involves multiple and possibly conflicting objectives. Algorithms designed to deal with a single objective can be highly unfair in terms of their rewards in the other objectives. To overcome this issue, we present the Pareto Contextual Zooming (PCZ) algorithm which achieves $\tilde{O}(T^{(1+d_p)/(2+d_p)})$ Pareto regret where d_p is the Pareto zooming dimension. 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.

Acknowledgments

This material is based upon work supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under 3501 Program Grant No. 116E229.

References

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

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

[3] 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.

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

[5] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 485–492, 2010.

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

[7] C. Tekin and E. Turgay, "Multi-objective contextual multi-armed bandit problem with a dominant objective," *arXiv preprint arXiv:1708.05655*, 2017.

[8] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, "Contextual bandits with linear payoff functions," in *Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

[9] M. M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *Proc. 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013.

[10] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 20, pp. 817–824, 2009.

[11] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang, "Efficient optimal learning for contextual bandits," in *Proc. 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 169–178, 2011.

[12] 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. International Conference on Machine Learning (ICML)*, pp. 1638–1646, 2014.

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

[14] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proc. International Conference on Machine Learning (ICML)*, pp. 757–765, 2014.

[15] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proc. 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 539–548, ACM, 2016.

[16] S. Q. Yahyaa, M. M. Drugan, and B. Manderick, "Annealing-Pareto multi-objective multi-armed bandit algorithm," in *Proc. 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2014.

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

[18] K. Van Moffaert, K. Van Vaerenbergh, P. Vrancx, and A. Nowé, "Multi-objective χ-armed bandits," in *Proc. 2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 2331–2338, 2014.

[19] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "χ-armed bandits," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1655–1695, 2011.