Ranked Prioritization of Groups in Combinatorial Bandit Allocation

Lily Xu, Arpita Biswas, Fei Fang and Milind Tambe

1 Center for Research on Computation and Society, Harvard University
2 Institute for Software Research, Carnegie Mellon University

Abstract

Preventing poaching through ranger patrols is critical for protecting endangered wildlife. Combinatorial bandits have been used to allocate limited patrol resources, but existing approaches overlook the fact that each location is home to multiple species in varying proportions, so a patrol benefits each species to differing degrees. When some species are more vulnerable, we ought to offer more protection to these animals; unfortunately, existing combinatorial bandit approaches do not offer a way to prioritize important species. To bridge this gap, (1) We propose a novel combinatorial bandit objective that trades off between reward maximization and also accounts for prioritization over species, which we call ranked prioritization. We show this objective can be expressed as a weighted linear sum of Lipschitz-continuous reward functions. (2) We provide RankedCUCB, an algorithm to select combinatorial actions that optimize our prioritization-based objective, and prove that it achieves asymptotic no-regret. (3) We demonstrate empirically that RankedCUCB leads to up to 38% improvement in outcomes for endangered species using real-world wildlife conservation data. Along with adapting to other challenges such as preventing illegal logging and overfishing, our no-regret algorithm addresses the general combinatorial bandit problem with a weighted linear objective.

1 Introduction

More than a quarter of mammals assessed by the IUCN Red List are threatened with extinction [Gilbert, 2008]. As part of the UN Sustainable Development Goals, Target 15.7 focuses on ending poaching, and Target 15.5 directs us to halt the loss of biodiversity and prevent the extinction of threatened species. To efficiently allocate limited resources, multilarmed bandits (MABs), and in particular combinatorial bandits [Chen et al., 2016; Cesa-Bianchi and Lugosi, 2012], have been widely used for a variety of tasks [Bastani and Bayati, 2020; Segal et al., 2018; Misra et al., 2019] including ranger patrols to prevent poaching [Xu et al., 2021a]. For poaching prevention, the patrol planner is tasked with repeatedly and efficiently allocating a limited number of patrol resources across different locations within the park [Plumptre et al., 2014; Fang et al., 2016].

In past work, we worked with the World Wide Fund for Nature (WWF) to deploy machine learning approaches for patrol planning in Srepok Wildlife Sanctuary in Cambodia [Xu et al., 2020; Xu et al., 2021b]. Subsequent conversations with park managers and conservation biologists raised the importance of protecting vulnerable species during patrols. For example, they pointed to the Uganda Wildlife Authority which ranked threatened species to allocate patrol effort in national parks, using the IUCN Red List as a key standard [Rodrigues et al., 2006]. In this paper, we address directly consider the conservation of vulnerable species to better align our algorithmic objectives with on-the-ground conservation priorities.

Looking at real-world animal distribution and snare locations in Srepok (Figure 1), we observe that the locations that maximize expected reward, defined as finding the most snares (darkest red in the risk map), are not necessarily the regions with high density of endangered animals (elephants). To effectively improve conservation outcomes, we must account for these disparate impacts, as the relatively abundant muntjac and wildpig would benefit most if we simply patrol the regions with most snares, neglecting the endangered elephant. Prioritization of species is well-grounded in existing conservation best practices [Regan et al., 2008; Arponen, 2012; Dilkina et al., 2017]. For example, Plumptre et al. [2014] propose planning ranger patrols by setting conservation targets for key species. We term the goal of preferentially allocating resources to maximize benefit to priority groups as ranked prioritization.

![Figure 1: Animal density distributions and poaching risk (expected number of snares) across a 10 × 10 km region in Srepok Wildlife Sanctuary, Cambodia. The distribution of the species of greatest conservation concern, the elephant, differs from that of other animals and overall poaching risk.](image-url)
Some existing multi-armed bandit approaches have considered priority-based fairness [Joseph et al., 2016; Kearns et al., 2017; Schumann et al., 2022]. However, these prior works only consider stochastic, not combinatorial, bandits. In our combinatorial bandit problem setup, we must determine some hours of ranger patrol to allocate across $N$ locations at each timestep, subject to a budget constraint. The rewards obtained from these actions are unknown a priori so must be learned in an online manner. Existing combinatorial bandit approaches [Chen et al., 2016; Xu et al., 2021a] achieve no-regret guarantees with respect to the objective of maximizing the direct sum of rewards obtained across all arms. However, we wish to directly consider ranked prioritization of groups in our objective; learning combinatorial rewards in an online fashion while ensuring we justly prioritize vulnerable groups along with reward makes the problem more challenging. How to trade off reward and prioritization in the combinatorial bandit setting has so far remained an open question. We show experimentally that straightforward solutions fail to make the appropriate trade-off.

To improve wildlife conservation outcomes by addressing the need for species prioritization, we contribute to the literature on online learning for resource allocation across groups with the following:

1. We introduce a novel problem formulation that considers resource allocation across groups with ranked prioritization, which has significant implications for wildlife conservation, as well as illegal logging and poaching. We show that an objective that considers both prioritization and reward can be expressed as a weighted linear sum, providing the useful result that the optimal action in hindsight can be found efficiently via a linear program.

2. We provide a no-regret learning algorithm, RankedCUCB, for the novel problem of selecting a combinatorial action (amount of patrol effort to allocate to each location) at each timestep to simultaneously attain high reward and group prioritization. We prove that RankedCUCB achieves sub-linear regret of $O\left(\frac{\ln T}{N} + \frac{N}{\Delta} \right)$ for a setting with $N$ locations, $O(J^N)$ combinatorial actions, and time horizon $T$.

3. Using real species density and poaching data from Srepok Wildlife Sanctuary in Cambodia along with synthetic data, we experimentally demonstrate that RankedCUCB achieves up to 38% improvement over existing approaches in ranked prioritization for endangered wildlife.

## 2 Problem Formulation

Consider a protected area with $N$ locations and $G$ groups of interest, with each group representing a species or a set of species in the context of poaching prevention. Let $d_{gr}$ denote the (known) fraction of animals of a group $g \in [G]$ present in location $i \in [N]$. Note that $\sum_{i=1}^{N} d_{gi} = 1$ for all $g \in [G]$. We assume the groups are disjoint, i.e., each animal is a member of exactly one group, and the impact of an action on a location equally benefits all animals that are present there.

At each timestep $t$ until the horizon $T$, the planner determines an action, which is an effort vector $\beta^{(t)} = (\beta_1, \ldots, \beta_N)$ specifying the amount of effort (e.g., number of patrol hours) to allocate to each location. The total effort at each timestep is constrained by a budget $B$ such that $\sum_{i=1}^{N} \beta_i \leq B$. To enable implementation in practice, we assume the effort is discretized into $J$ levels, thus $\beta_i \in \Psi = \{\psi_1, \ldots, \psi_J\}$ for all $i$.

After taking action $\beta$, the decision maker receives some reward $\mu(\beta)$. We assume the reward is decomposable [Chen et al., 2016], defined as the sum of the reward at each location:

$$\text{expected reward} = \mu(\beta) = \sum_{i=1}^{N} \mu_i(\beta_i).$$

We assume $\mu_i \in [0, 1]$ for all $i$. In the poaching context, the reward $\mu_i$ corresponds to the true probability of detecting snares at a location $i$. The reward function is initially unknown, leading to a learning problem. Following the model of combinatorial bandit allocation from Xu et al. [2021a], we assume that the reward function $\mu_i(\cdot)$ is (1) Lipschitz continuous, i.e., $|\mu_i(\psi_j) - \mu_i(\psi_k)| \leq \Delta |\psi_j - \psi_k|$ for some Lipschitz constant $\Delta$ and all pairs $\psi_j, \psi_k$, and (2) monotonically non-decreasing such that exerting greater effort never results in a decrease in reward, i.e., $\beta'_i > \beta_i$ implies $\mu_i(\beta'_i) \geq \mu_i(\beta_i)$.

Finally, we assume that an ordinal ranking is known to indicate which group is more vulnerable. The ranking informs the planner which groups are of relatively greater concern, as we often lack sufficient data to quantify the extent to which one group should be prioritized. Without loss of generality, assume the groups are numerically ordered with $g = 1$ being the most vulnerable group and $g = G$ being the least, so the true rank is $\text{rank} = (1, \ldots, G)$.

### 2.1 Measuring Ranked Prioritization

To evaluate prioritization, we must measure how closely the outcome of a combinatorial action $\beta$ aligns with the priority ranking over groups. Later in Section 6, we discuss how our approach can generalize to other prioritization definitions, such as the case where have greater specificity over the relative prioritization of each group.

To formalize prioritization, we define $\text{benefit}(\beta; g)$ which quantifies the benefit group $g$ receives from action $\beta$. As mentioned earlier, we assume that reward $\mu_i(\beta_i)$ obtained by taking an action $\beta_i$ impacts all individuals at location $i$. Let $n_{gi}$ be the number of individuals from group $g$ at location $i$ and $n_g$ be the total number of individuals in group $g$. We define

$$\text{benefit}(\beta; g) := \frac{\sum_{i=1}^{N} n_{gi} \mu_i(\beta_i)}{n_g}.\quad (2)$$

An action $\beta$ perfectly follows ranked prioritization when

$$\text{benefit}(\beta; 1) \geq \text{benefit}(\beta; 2) \geq \cdots \geq \text{benefit}(\beta; G).$$
However, when \( \tilde{\beta} \) does not follow perfect rank prioritization, we need a metric to measure the extent to which groups are appropriately prioritized. This metric would quantify similarity between the ranking induced by the benefits of \( \tilde{\beta} \) and the true rank. One common approach to evaluate the similarity between rankings is the Kendall \( \tau \) coefficient:

\[
\text{Kendall } \tau = \frac{\text{# concordant pairs} - \text{# discordant pairs}}{\binom{G}{2}} .
\]

(3)

The number of concordant pairs is the number of pairwise rankings that agree, across all \( \binom{G}{2} \) pairwise rankings; the discordant pairs are those that disagree. This metric yields a value \([-1, 1]\) which signals amount of agreement, where \(+1\) is perfect agreement and \(-1\) is complete disagreement (i.e., reverse ordering). However, a critical weakness of Kendall \( \tau \) is that it is discontinuous, abruptly jumping in value when the effort \( \beta_i \) on an arm is perturbed, rendering optimization difficult.

Fortunately, we have at our disposal not just each pairwise ranking, but also the magnitude by which each pair is in concordance or discordance. Leveraging this information, we construct a convex function \( \mathcal{P}(\tilde{\beta}) \) that quantifies group prioritization:

\[
\mathcal{P}(\tilde{\beta}) = \sum_{g,h \in [G]} \mathbb{I}(g < h) \cdot \left( \text{benefit}(\tilde{\beta}; g) - \text{benefit}(\tilde{\beta}; h) \right) \binom{G}{2}
\]

where \( \mathbb{I}(g < h) = 1 \) if \( g \) is of higher priority than \( h \) according to the true rank, and 0 otherwise. When \( g < h \), the difference in benefit to group \( g \) and \( h \) is positive if the pairs are concordant and negative if the pairs are discordant. Intuitively, for each concordant pair, we earn \( (\text{benefit}(\tilde{\beta}; g) - \text{benefit}(\tilde{\beta}; h)) \), and for each discordant pair, we are penalized this quantity. The value \( \mathcal{P}(\tilde{\beta}) \) can be interpreted as the correlation between the priority outcomes of \( \tilde{\beta} \) and the true priority ranking, giving us a metric for evaluating the group prioritization of an action \( \tilde{\beta} \).

Our prioritization metric \( \mathcal{P}(\tilde{\beta}) \) induces nice properties. First, this metric is robust to perturbations in \( \mu_i(\beta) \) values whereas Kendall \( \tau \) is not. As \( \mathcal{P}(\cdot) \) is continuous and differentiable, we can use it directly in the objective function. Second, with slight modifications to \( \mathcal{P} \) we can seamlessly adapt to other useful settings, as we discuss in Section 6.

Leveraging the fixed ordering \( \text{rank} = (1, \ldots, G) \) over groups, we can simplify Equation (4) as

\[
\mathcal{P}(\tilde{\beta}) = \sum_{g=1}^{G-1} \sum_{h=g+1}^{G} \left( \text{benefit}(\tilde{\beta}; g) - \text{benefit}(\tilde{\beta}; h) \right) \binom{G}{2}
\]

\[
= \sum_{g=1}^{G-1} \sum_{h=g+1}^{G} \left( \sum_{i=1}^{N} \mu_i(\beta) d_{gi} - \sum_{i=1}^{N} \mu_i(\beta) d_{hi} \right) \binom{G}{2}
\]

\[
= \sum_{i=1}^{N} \mu_i(\beta) \sum_{g=1}^{G-1} \sum_{h=g+1}^{G} (d_{gi} - d_{hi}) \binom{G}{2} .
\]

(5)

### 2.2 Objective with Prioritization and Reward

We wish to take actions that maximize our expected reward (Eq. 1) while also distributing our effort across the various groups as effectively as possible (Eq. 5). Recognizing that targeting group prioritization requires us to sacrifice reward, we set up the objective to balance reward and prioritization with a parameter \( \lambda \in (0, 1] \) that enables us to tune the degree to which we emphasize reward vs. prioritization:

\[
\text{obj}(\tilde{\beta}) = \lambda \mu(\tilde{\beta}) + (1 - \lambda) \cdot \mathcal{P}(\tilde{\beta}) .
\]

(6)

### 3 Approach

We show that the objective (Eq. 6) can be reformulated as a weighted linear combination of the reward of each individual arm, which we then solve using our RankedCUCB algorithm, producing a general-form solution for combinatorial bandits.

Using the prioritization metric from Equation (5), we can re-express the objective (6) as

\[
\text{obj}(\tilde{\beta}) = \sum_{i=1}^{N} \mu_i(\beta) \left( \lambda + (1 - \lambda) \frac{\sum_{g=1}^{G-1} \sum_{h=g+1}^{G} (d_{gi} - d_{hi}) \binom{G}{2}}{\binom{G}{2}} \right)
\]

(7)

Observe that the large parenthetical is comprised only of known constants, so we can precompute the parenthetical value as \( \Gamma_i \) for each location \( i \) to yield

\[
\text{obj}(\tilde{\beta}) = \sum_{i=1}^{N} \mu_i(\beta) \Gamma_i ,
\]

(8)

providing the useful result that our objective is decomposable across the locations \( i \in [N] \) — specifically that the objective is a linear combination of the reward for each component of the action.

### 3.1 Combinatorial UCB Approach

The expected reward functions \( \mu_i(\cdot) \) are unknown in advance and stochastic, so our setting falls within the space of sequential learning problems that need to be learned over multiple timesteps. Therefore, an action (an effort vector \( \tilde{\beta} \)) we take in each timestep can be viewed as an arm pull under a multi-armed bandit framework. Specifically, we use a combinatorial bandit approach [Chen et al., 2016], where each arm is an effort level \( \psi_j \) at one location \( i \), rendering \( N J \) total arms. We track the reward of each arm separately then use an oracle to select a combined action \( \tilde{\beta} \), which is a combinatorial selection of arms. Our goal is to design an algorithm that determines an action at each time \( t \) to minimize regret, defined as the difference between the total expected reward of an optimal action and the reward achieved following actions recommended by the algorithm.

To balance exploration (learning the reward of actions with greater uncertainty) with exploitation (taking actions known to be high-reward), we employ an upper confidence bound (UCB) approach that considers for each action its estimated reward along with our uncertainty over that estimate. Let \( \hat{\mu}_t(i,j) = \hat{\text{reward}}_t(i,\psi_j)/\hat{\mu}_t(i,\psi_j) \) be the average observed reward of location \( i \) at effort \( \psi_j \) given cumulative empirical
As the prioritization metric can be expressed as a linearly nondecreasing function, we can select a combinatorial action leveraged by the assumptions described in Section 2 that precisely estimate rank order.

We now prove that our iterative algorithm RankedCUCB (Alg. 1) guarantees no regret with respect to the optimal solution for the objective (8) that jointly considers reward and prioritization. More formally, we show that $\text{Regret}_{T} \to 0$ as $T \to \infty$ where

$$\text{Regret}_{T} := \mu(\hat{\beta}^*) - \frac{1}{T} \sum_{t=1}^{T} \mu(\hat{\beta}(t)) \cdot$$

Here, $\hat{\beta}^*$ is an optimal action and expected reward $\mu(\hat{\beta}) := \sum_{i=1}^{N} \mu_{i}(\hat{\beta}) \Gamma_{i}$ for an effort vector $\hat{\beta} = \{\beta_{1}, \ldots, \beta_{N}\}$. Note that if $\Gamma_{i} < 0$, any solution to the maximization problem (8) would allocate $\beta_{i} = 0$, and so would RankedCUCB. Hence, for the analysis we assume that we consider only those locations whose $\Gamma_{i} > 0$.

### 4.1 Convergence of Estimates

To prove the no-regret guarantee of RankedCUCB, we first establish Lemma 1 which states that, with high probability, the UCB estimate $\mu_{i}(i, j) \Gamma_{i} + r_{i}(i, j)$ converges within a confidence bound of $r_{i}(i, j) = \sqrt{3\Gamma_{i}^{2} \log T} / (2n_{i})$.

**Lemma 1.** Using RankedCUCB, after $t$ timesteps, each $\mu_{i}(i, j) \Gamma_{i}$ estimate converges to a value within a radius $r_{i}(i, j) = \sqrt{3\Gamma_{i}^{2} \ln t} / (2n_{i}(i, j))$ of the corresponding true $\mu_{i}(i, j) \Gamma_{i}$ value with probability $1 - \frac{2N^{2} J}{t}$ for all $i, j$.

**Proof Sketch.** Using the Chernoff–Hoeffding bound, we show that, at any timestep $t$, the probability that the difference between $\mu_{i}(i, j) \Gamma_{i}$ and $\mu_{i}(i, j) \Gamma_{i}$ is greater than $r_{i}(i, j)$, is at most $2/t^5$. We then use union bound to show that $\mu_{i}(i, j) \Gamma_{i}$ converges to a value within radius $r_{i}(i, j)$ of $\mu_{i}(i, j) \Gamma_{i}$ with probability $1 - \frac{2N^{2} J}{t}$. The complete proof is given in Appendix B.2.

### 4.2 Achieving No Regret

**Theorem 1.** The cumulative regret of RankedCUCB is $O \left( \frac{2N^{2}}{T} + NJ \right)$ with $N$ arms, $J$ discrete effort values, and time horizon $T$.

---

**Algorithm 1 RankedCUCB**

**Input:** time horizon $T$, budget $B$, discretization levels $\Psi = \{\psi_{1}, \ldots, \psi_{J}\}$, arms $i \in [N]$ with unknown reward $\mu_{i}$

**Parameters:** tuning parameter $\lambda$

1: Precompute $\Gamma_{i}$ for each arm $i \in [N]$
2: $n(i, \psi_{j}) = 0, \text{reward}(i, \psi_{j}) = 0 \quad \forall i \in [N], j \in [J]$
3: for timestep $t = 1, 2, \ldots, T$ do
4:  Let $\varepsilon_{t} = t^{-1/3}$
5:  Compute $\text{UCB}_{i}$ for all arms using Eq. (11)
6:  Solve $\text{LP}(\text{UCB}_{i},\{\Gamma_{i}\}, B)$ to select super arm $\hat{\beta}$
7:  // Execute action
8:  Act on $\hat{\beta}$ to observe rewards $X_{i}^{t}, X_{2}^{t}, \ldots, X_{N}^{t}$
9:  for arm $i = 1, \ldots, N$ do
10:    $\text{reward}(i, \beta_{i}) = \text{reward}(i, \beta_{i}^{*}) + X_{i}^{t}$
11:    $n(i, \beta_{i}) = n(i, \beta_{i}) + 1$

---
RankedCUCB performs significantly better than baselines. Hence the expected regret is $O(\ln T/N + NJ)$. The complete proof is given in Appendix B.3.

Our analysis improves upon the regret analysis for standard (non-weighted) combinatorial bandits [Chen et al., 2016] by not having to rely on the “bounded smoothness” assumption of reward functions. The technical tools we employ for this regret analysis can be of independent interest to the general multi-armed bandits community.

### 5 Experiments

We evaluate the performance of our RankedCUCB approach in real-world conservation data from Srepok Wildlife Sanctuary to demonstrate significant improvement over approaches that ignore prioritization.\(^1\)

Of the key species in Srepok on which we have species density data, we prioritize elephants, followed by banteng, muntjac, and wild pig, corresponding to their IUCN Red List categories of critically endangered, endangered, least concern but decreasing, and least concern respectively. Thus our setting has $G = 4$ groups distributed across $N = 25$ locations within the park. We further evaluate our algorithm on a synthetic dataset with $G = 5$ groups randomly distributed across the park.

#### Baselines

We compare the performance of our RankedCUCB approach with naive priority-aware, priority-blind, random, and optimal benchmarks. NaiveRank takes the straightforward approach of directly solving for the objective that weighs each target by its individual ranked prioritarian value, which accounts for prioritization induced by each target independently but ignores the coupled effect across targets. LIZARD [Xu et al., 2021a] maximizes the combinatorial reward objective; this algorithm enhances standard UCB approaches by accounting for smoothness and monotonicity in the reward function, but ignores prioritization. Optimal solves for the action $\hat{\beta}^* = \arg \max \theta \mathbf{obj}(\theta)$ that maximizes the objective, using the ground truth values of $\mu_i$ to directly compute the best action. Random selects an arbitrary subset of actions that satisfies the budget constraint.

See Figure 2 for a comparison of the approaches. RankedCUCB outperforms all baselines, and the breakdown of the reward and prioritization components reveals that this gain is a result of the tradeoff between the two components: although LIZARD is able to learn high-reward actions, these actions lead to prioritization outcomes that are worse than random. LIZARD even achieves reward that exceeds that of the optimal action (which also considers rank priority), but that comes at the cost of poor performance.

---

\(^1\) Code is available at https://github.com/lily-x/rankedCUCB
prioritization. Notably, NaiveRank performs worse than random, even measured on fairness: focusing on the individual targets with the best prioritization neglects group-wide patterns throughout the park.

The reward–prioritization tradeoff becomes more apparent when we analyze the Pareto frontier in Figure 2(right), which plots the reward and prioritization of each approach as we change the value of $\lambda$. The price of prioritization is clear: the more we enforce prioritization (smaller values of $\lambda$) the lower our possible reward. However, this tradeoff is not one-for-one; the steepness of the Pareto curve allows us to strategically select the value of $\lambda$. For example, we achieve a nearly two-fold increase in prioritization by going from $\lambda = 0.9$ to 0.7 while sacrificing less than 10% of the reward, indicating a worthwhile tradeoff.

6 Generalizability to Other Settings

Beyond wildlife conservation, our problem formulation applies to domains with the following characteristics: (1) repeated resource allocation, (2) multiple groups with some priority ordering, (3) an a priori unknown reward function that must be learned over time, and (4) actions that impact some subset of the groups to differing degrees. Our approach also adapts to the following related objectives.

**Weighted rank.** Gatmiry et al. [2021] suggest some specific metrics for how rangers should prioritize different species, noting ranger enforcement of poaching urial (a wild sheep) should be 2.5 stricter than red deer. Our approach could easily adapt to these settings where domain experts have more precise specification of desired outcomes. In this case, we introduce a parameter $\alpha_g$ for each group $g$ that specifies the relative importance of group $g$, leading to the following prioritization metric:

$$\mathcal{P}_{\text{CF}}(\vec{\beta}) = \sum_{i=1}^{N} \mu_i(\beta_i) \sum_{g=1}^{G-1} \sum_{h=g+1}^{G} \alpha_g d_{gi} - \alpha_h d_{hi}$$

where we set $\alpha_g > \alpha_h$ when $g < h$ to raise our threshold of how strongly group $g$ should be favored. From our example, we would set $\alpha_{\text{urial}} = 2.5 \alpha_{\text{deer}}$. This metric aligns with the definition of calibrated fairness [Liu et al., 2017].

**Weighted reward.** We can also accommodate the setting where the reward is a weighted combination, i.e., if reward is $N \sum_{i=1}^{N} c_i \mu_i(\beta_i)$ for some coefficients $c_i \in \mathbb{R}$, as the coefficients $c_i$ would be absorbed into the $\Gamma_i$ term of the objective.

7 Related Work

**Multi-armed bandits.** MABs [Lattimore and Szepesvári, 2020] have been applied to resource allocation for healthcare [Bastani and Bayati, 2020], education [Segal et al., 2018], and dynamic pricing [Misra et al., 2019]. These papers solve various versions of the stochastic MAB problem [Auer et al., 2002].

Several prior works consider resource allocation settings where each arm pull is costly, limiting the total number of pulls by a budget. Tran-Thanh et al. [2010] use an $\varepsilon$-based approach to achieve regret linear in the budget $B$; Tran-Thanh et al. [2012] provide a knapsack-based policy using UCB to improve to $O(\ln B)$ regret. Chen et al. [2016] extend UCB1 to a combinatorial setting, matching the regret bound of $O(\log T)$. Slivkins [2013] address ad allocation where each stochastic arm is limited by a budget constraint on the number of pulls. Our objective (8) is similar to theirs but they do not handle combinatorial actions, which makes our regret analysis even harder.

**Prioritization notions in MABs.** Addressing prioritization requires reward trade-offs, which is related to objectives from algorithmic fairness [Dwork et al., 2012; Kleinberg et al., 2018; Corbett-Davies et al., 2017]. However, most MAB literature favor traditional reward-maximizing approaches [Lattimore et al., 2015; Agrawal and Devanur, 2014; Verma et al., 2019]. The closest metric to our prioritization objective is meritocratic fairness [Kearns et al., 2017]. For MAB, meritocratic fairness was introduced by Joseph et al. [2016], to ensure that lower-reward arms are never favored above a high-reward one. Liu et al. [2017] apply calibrated fairness by enforcing a smoothness constraint on our selection of similar arms, using Thompson sampling to achieve an $O(T^{2/3})$ fairness regret bound. For contextual bandits, Chohlas-Wood et al. [2021] take a consequentialist (outcome-oriented) approach and evaluate possible outcomes on a Pareto frontier, which we also investigate. Wang et al. [2021] ensure that arms are proportionally represented according to their merit, a metric they call fairness of exposure. Others have considered fairness in multi-agent bandits [Hossain et al., 2021], restless bandits [Herlihy et al., 2021], sleeping bandits [Li et al., 2019], contextual bandits with biased feedback [Schumann et al., 2022], and infinite bandits [Joseph et al., 2018]. However, none of these papers consider correlated groups, which is our focus.

8 Conclusion

We address the challenge of allocating limited resources to prioritize endangered species within protected areas where true snare distributions are unknown, closing a key gap identified by our conservation partners at Srepok Wildlife Sanctuary. Our novel problem formulation introduces the metric of ranked prioritization to measure impact across disparate groups. Our RankedCUCB algorithm offers a principled way of balancing the tradeoff between reward and ranked prioritization in an online learning setting. Notably, our theoretical guarantee bounding the regret of RankedCUCB applies to a broad class of combinatorial bandit problems with a weighted linear objective.

Acknowledgments

This work was supported in part by NSF grant IIS-2046640 (CAREER) and the Army Research Office (MURI W911NF-18-1-0208). Biswas was supported by the Center for Research on Computation and Society (CRCS). Thank you to Andrew Plumptre for a helpful discussion on species prioritization in wildlife conservation; Jessie Finocchiaro for comments on an earlier draft; and all the rangers on the front lines of biodiversity efforts in Srepok and around the world.

5211
References


