Stochastic Bandits for Egalitarian Assignment

Anonymous authors
Paper under double-blind review

Abstract

We study `EgalMAB`, an egalitarian assignment problem in the context of stochastic multi-armed bandits. In `EgalMAB`, an agent is tasked with assigning a set of users to arms. At each time step, the agent must assign exactly one arm to each user such that no two users are assigned to the same arm. Subsequently, each user obtains a reward drawn from the unknown reward distribution associated with its assigned arm. The agent's objective is to maximize the minimum expected cumulative reward among all users over a fixed horizon. This problem has applications in areas such as fairness in job and resource allocations, among others. We design and analyze a UCB-based policy `EgalUCB` and establish upper bounds on the cumulative regret. In complement, we establish an almost-matching policy-independent impossibility result.

1 Introduction

The multi-armed bandit (MAB) problem serves as a model for online decision-making under uncertainty, finding applications in diverse domains (Shen et al., 2015; Durand et al., 2018; Ding et al., 2019; Mueller et al., 2019; Forouzandeh et al., 2021). In the classical stochastic MAB problem, an agent is provided with a set of K arms, each associated with an unknown distribution. At each round, the agent plays an arm and receive a reward drawn from its distribution. The agent's goal is to maximize the expected cumulative reward obtained over a fixed number of time steps T.

In our work, we study the problem of *egalitarian assignment* in the context of stochastic MABs, which we refer to as `EgalMAB`. In this scenario, the agent is provided with a set of $U < K$ users. At each time step, the agent must assign exactly one arm to each user such that no two users are assigned to the same arm. Subsequently, each user obtain a reward drawn from the reward distribution associated with its assigned arm. The agent's objective is to *maximize the minimum* expected cumulative reward among all users.

`EgalMAB` finds applications in various domains, including ensuring fairness in job and resource allocations. Consider the job assignment problem depicted in Figure 1. In this scenario, there are U users with recurring jobs of equal load (e.g., hourly database updates) and K shared cloud computing resources with fluctuating computational power (e.g., due to unrelated loads that are running on neighbouring cores of the same physical node (Kousiouris et al., 2011)). The agent's objective is to distribute these jobs *fairly* among the available resources such that over T trials, the user who has to wait the longest across all their jobs (i.e., receives the least cumulative reward) is not significantly worse off compared to other users.

Similarly, ride-hailing services present another scenario where fair assignment is desirable. In this scenario, there are U passengers and K drivers who offer varying degrees of experience to passengers (e.g., due to vehicle condition and driver behavior). The agent's objective is to allocate users *fairly* to vehicles such that over T trips, no passenger faces a significantly worse overall experience than others.

As illustrated in both examples, `EgalMAB` embodies the principle of egalitarianism, which is also known as the Rawlsian maximin principle (Rawls, 1971), in its notion of fairness. Egalitarianism is a fundamental notion of justice based on the *difference principle*, which seeks to *maximize* the welfare of those in society who are the *worst-off*. Likewise, our agent's objective is to maximize the cumulative reward of the user receiving the lowest cumulative reward among all U users. Given our examples above, solving the `EgalMAB` problem yields a policy that minimizes the overall waiting time of the user who has to wait the longest and

Figure 1: An illustration with $K = 8$ arms and $U = 7$ users. After time step T, each user $u \in [U]$ has some expected cumulative reward $\mathbb{E}[S_{u,T}]$. The agent's objective is to maximize $\min_{u\in[U]} \mathbb{E}[S_{u,T}]$, which is the minimum expected cumulative reward across all users.

maximizes the experience of the passenger who has the most negative encounters. This is in contrast to the classic MAB setting where only *overall utility* is considered. This prompts the fundamental question that our work seeks to answer:

> *How can we design an agent's assignment policy that optimizes overall utility for individual users while also ensuring that no user within a group consistently encounters substantially sub-optimal outcomes?*

Our contributions are summarized as follows. We formally define the `EgalMAB` problem in Section 3 and propose `EgalUCB`, a UCB-based solution to `EgalMAB`, in Section 4. In Section 5 and 6, we establish that `EgalUCB` achieves an expected regret of at most

$$O\left(\sqrt{\frac{T\ln(T)\cdot(K-U)}{U}}\right).$$

We also provide a policy-independent lower bound that matches the upper bound up to a multiplicative factor of $\sqrt{1/U}$ and a term logarithmic in T. Proof sketches for these results are included in Section 6. Lastly, empirical validations for these results using both synthetic and real-world data are presented in Section 7.

2 Related Works

MAB with Multiple Plays. The multiple-play multi-armed bandit (MP-MAB) problem (Anantharam et al., 1987; Gai et al., 2012; Chen et al., 2016; Komiyama et al., 2015) expands upon the classical MAB framework by allowing the agent to play $U \geq 2$ distinct arms. The MP-MAB problem has been extended in various ways, including combinatorial bandits (Cesa-Bianchi & Lugosi, 2012; Kveton et al., 2015a; Chen et al., 2016), cascading bandits (Kveton et al., 2015b; Wen et al., 2017), and MP-MAB with shareable arms (Wang et al., 2022). Although similar to `EgalMAB` in that the agent has to select multiple arms, unlike `EgalMAB`, these formulations of the MP-MAB problem do not involve *multiple users*; that is, the reward is with respect to the agent instead of users.

MAB with Multiple Users. The extension of multiple users into the classic MAB problem has been extensively explored in the context of cognitive radio networks (Jouini et al., 2009; Liu & Zhao, 2010; Avner & Mannor, 2014). Unlike `EgalMAB` where a *centralized* agent assigns the arms to the users, these works have focused on scenarios where multiple *decentralized* users interact with a single MAB instance. As a result of this decentralization, it is possible for multiple users to play the same arm simultaneously, resulting in a collision that can negatively affect the received reward.

Fairness in MAB. The introduction of fairness considerations into the classical MAB problem has garnered significant interest. Much attention has been directed towards addressing fairness concerns that

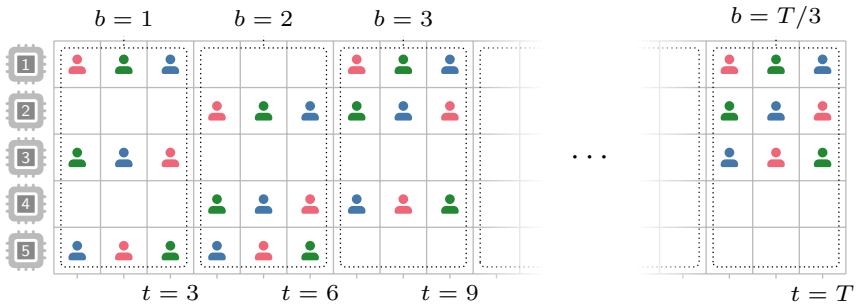

Figure 2: A trace of `EgalUCB` with $K = 5$ arms and $U = 3$ users. When $b = 2$, the three arms with the highest $\mathrm{UCB}_{a,b}$ values are $a \in \{2, 4, 5\}$, which are then assigned in a round-robin fashion across time steps $t \in \{4, 5, 6\}$. As b increases, the estimates for μ_a for all arms a improves. By $b = T/3$ for large T, the three arms with the highest $\mathrm{UCB}_{a,b}$ values are most likely $a \in \{1, 2, 3\}$, which are then assigned over time steps $t \in \{T-2, T-1, T\}$.

typically involve ensuring that each arm is played a minimum number of times, known as fairness in exposure (Claure et al., 2020; Chen et al., 2020; Li et al., 2020; Wang et al., 2021), or with a probability proportional to the arm's merit, known as meritocratic fairness (Joseph et al., 2016; 2018). While `EgalMAB` emphasizes fairness among users, the aforementioned works deal with fairness among arms. However, there are alternative approaches that maintain a focus on fairness among users. One such approach involves maximizing the Nash social welfare (NSW) function. This function is defined as the product of rewards obtained by all users, which intuitively encodes the notion of fairness as it increases only when most users achieve high rewards. Hossain et al. (2021) investigated a scenario involving U users and K arms, in which each arm's reward varies for each user due to different perceived utilities. In each time step, only one arm is played, and all users obtain rewards according to their perceived utilities. Their primary objective is to maximize the NSW. Sawarni et al. (2023) examined an alternative fairness framework within the context of linear bandits. In their scenario, a new user is introduced at each time step, and fairness is ensured by considering the product of rewards across all time steps. Both of these works assume that only a *single* arm is played at each time step, while the agent in `EgalMAB` simultaneously selects *multiple* arms. Additionally, instead of aiming to maximize the *NSW function*, `EgalMAB` focuses on maximizing the *cumulative reward of the worst-off user*.

3 `EgalMAB` Problem

In this section, we will formally define the components involved in an `EgalMAB` problem. We will start by presenting a working definition for each component. When relevant, we will supplement these definitions with the measure-theoretic details necessary for the proofs of lower bounds.

Environment. Let K be the number of arms, $U < K$ be the number of users, and T be the time horizon. For each arm $a \in [K]$, let p_a denote the reward density. An instance of the `EgalMAB` problem is represented by the tuple (ν, U, T) where $\nu \coloneqq (p_1, \ldots, p_K)$. When the context is clear, we also refer to ν as an `EgalMAB` instance. We assume that the expected reward $\mu_a \coloneqq \mathbb{E}_{X \sim p_a}[X]$ obtained for playing arm a is finite. Moreover, for convenience, we assume that the arms are indexed in such a way that $\mu_1 \geq \cdots \geq \mu_K$. However, the agent is unaware of this ordering.

More formally, for each arm $a \in [K]$, let $\mathbb{P}_a$ be the probability law for the reward obtained after playing arm a. For any $\mathbb{P}_a$-measurable set B, we have $\mathbb{P}_a(B) = \int_B p_a \, \mathrm{d}\lambda$, where λ is the Lebesgue measure and $p_a = \mathrm{d}\mathbb{P}_a / \mathrm{d}\lambda$ is the Radon–Nikodym derivative for $\mathbb{P}_a$.

Agent Policy. A solution to the `EgalMAB` problem is characterized by an agent policy $\pi \coloneqq (\pi_t)_t$. During each time step $t \in [T]$, the map π_t considers the actions and rewards history (which we will define shortly, after

introducing relevant notations) and assigns each user $u \in [U]$ to an arm $A_{u,t} \in [K]$ such that no two users are assigned the same arm. Subsequently, each user u receives a reward $X_{u,t}$ drawn independently from the density $p_{A_{u,t}}$. Using these notations, we will denote the history that π_t considers as $(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ where $A_t \coloneqq (A_{u,t})_u$ and $X_t \coloneqq (X_{u,t})_u$.

More formally, let $\mathcal{P}(S)$ denote the power set of a set S. For each time step $t \in [T]$, the stochastic map

$$\pi_t : (([K] \times \mathbb{R})^U)^{t-1} \times \mathcal{P}([K]^U) \to [0, 1]$$

is a probability kernel. We denote $\mathbb{P}_{\pi\nu}$ as the probability law for the interaction between the policy π and the `EgalMAB` instance ν over T time steps. Thus, the density of $\mathbb{P}_{\pi\nu}$ is

$$p_{\pi\nu}(A_1, X_1, \ldots, A_T, X_T) = \prod_{t=1}^{T} \pi_t(A_t | A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \prod_{u=1}^{U} p_{A_{u,t}}(X_{u,t}).$$

Egalitarian Objective. To achieve egalitarian fairness, we want to design a policy π that maximizes the expected cumulative reward of the least-rewarded user. Let

$$S_{u,t} \coloneqq \sum_{s=1}^{t} X_{u,s}$$

be the cumulative reward for user u up to time t. Formally, our egalitarian objective is to maximize $\min_{u \in [U]} \mathbb{E}[S_{u,T}]$. However, in line with most works in MAB, we will frame our problem as regret minimization. We define the *expected cumulative regret* as

$$R_T \coloneqq \frac{T\mu_*}{U} - \min_{u \in [U]} \mathbb{E}[S_{u,T}]$$

where $\mu_* \coloneqq \mu_1 + \cdots + \mu_U$ is the sum of the expected reward of the top U arms. The choice to compare with $T\mu_*/U$ is natural because the maximum expected cumulative reward obtained by the user with the least reward is at most $T\mu_*/U$. To see this, observe that the sum of cumulative rewards for all U users is $\sum_u \mathbb{E}[S_{u,T}] \leq T\mu_*$. Suppose, to the contrary, that $\min_u \mathbb{E}[S_{u,T}] > T\mu_*/U$, then $\sum_u \mathbb{E}[S_{u,T}] > T\mu_*$, which is a contradiction.

4 `EgalUCB` Policy

In this section, we describe our policy `EgalUCB` that achieves near-optimal regret for the `EgalMAB` problem. The `EgalUCB` policy is based on the `UCB1` policy (Auer et al., 2002) for solving classic MAB problems; however, it differs from the `UCB1` in several key aspects. We present the pseudocode for `EgalUCB` in Algorithm 1, complemented by a visual guide in Figure 2.

`EgalUCB` partitions the horizon into $B \coloneqq T/U$ blocks, each with U steps. We assume, without loss of generality, that T is divisible by U. In cases where this is not true, the difference in expected regret between the best and worst user is at most $(\mu_1 + \cdots + \mu_{\lceil U/2 \rceil}) - (\mu_{K-\lfloor U/2 \rfloor} + \cdots + \mu_U)$, which is independent of T.

Let $A \subseteq_U S$ denote that set A is a subset of set S with size U. `EgalUCB` begins by initializing some statistics (Lines 1–7). At the start of each block $b \in [B]$, it selects any set $A_b \subseteq_U [K]$ consisting of the highest-ranked U distinct arms as determined by their upper confidence bounds $\text{UCB}_{a,b}$ (Line 10). Over a block with U steps, these arms are then assigned to the U users in a round-robin fashion (Lines 11–23). After observing the rewards, `EgalUCB` updates its statistics (Lines 24–29).

Since each user is assigned to every arm in A_b exactly once, we have $\mathbb{E}[S_{u,T}] = \mathbb{E}[S_{u',T}]$ for all $u, u' \in [U]$. From a technical standpoint, this simplifies the regret analysis by eliminating the min operator in the regret R_T:

$$R_T = \frac{T\mu_*}{U} - \min_{u \in [U]} \mathbb{E}[S_{u,T}] = \frac{T\mu_*}{U} - \mathbb{E}[S_{u,T}], \quad \forall u \in [U].$$

Algorithm 1: `EgalUCB`

initialize current block $b = 0$
initialize current time step $t = 0$
foreach $a \in [K]$ **do**
  let number of blocks $B_{a,b} = 0$
  let cumulative reward $S_{a,t} = 0$
  let upper confidence bound $\text{UCB}_{a,b} = \infty$
end
while $b \leq T/U$ **do**
  update $b = b + 1$
  let $A_b \subseteq_U [K]$ be a set of U arms with highest $\text{UCB}_{a,b-1}$
  let $\text{ind} = (1, \ldots, U)$
  foreach $i \in [U]$ **do**
    update $t = t + 1$
    foreach $u \in [U]$ **do**
      let $A_{u,t} = A_b[\text{ind}[u]]$
    end
    play $(A_{1,t}, \ldots, A_{U,t})$ and receive $(X_{1,t}, \ldots, X_{U,t})$
    foreach $u \in [U]$ **do**
      let $a = A_b[\text{ind}[u]]$
      let $S_{a,t} = S_{a,t-1} + X_{u,t}$
    end
    circular shift ind by one to the right
  end
  foreach $a \in A_b$ **do**
    let $B_{a,b} = B_{a,b-1} + 1$
  end
  foreach $a \in [K]$ **do**
    let $\text{UCB}_{a,b} = \dfrac{S_{a,t}}{B_{a,b}U} + \sqrt{\dfrac{6\ln(bU)}{B_{a,b}U}}$
  end
end

5 Main Results

In this section, we present our main theoretical results. We first state some necessary definitions and notations. Then, we present the regret upper bounds for the `EgalUCB` policy. Following that, we discuss the policy-independent regret lower bound for `EgalMAB`.

Let $X_{a,t}$ be the reward obtained from playing arm a for the t-th time across all users. Note that if user u plays arm a at time step t, then the reward obtained is $X_{u,t} = X_{a,T_{a,t}}$ where $T_{a,t}$ is the number of times arm a is played up till time step t.

Let $B_{a,b}$ be the number of blocks that arm a is played up till block b. Note that $B_{a,b} = T_{a,bU}$. Furthermore, let

$$\hat{\mu}_{a,b} \coloneqq \frac{1}{bU} \sum_{t=1}^{bU} X_{a,t}$$

be the empirical estimate of μ_a after playing arm a for b blocks. As the policy observe more rewards, it gains confidence about its estimate of μ_a. This level of confidence is captured by

$$\epsilon_{b,b'} \coloneqq \sqrt{\frac{6\ln(bU)}{b'U}},$$

which is the confidence radius of playing an arm for b' blocks after block b.

Denote $A_* \coloneqq [U] \subseteq_U [K]$ as the set of best U arms. For any set of arms $A \subseteq [K]$ of not necessarily size U, let $A^- \coloneqq A \backslash A_*$ be the set A with the best U arms removed. Furthermore, let the expected reward of A be

$$\mu_A \coloneqq \sum_{a \in A} \mu_a.$$

The sub-optimality gap of A is defined as $\Delta_A \coloneqq \mu_{A_*} - \mu_A$. In the extreme cases, the maximum sub-optimality gap

$$\Delta_{\max} \coloneqq (\mu_1 + \cdots + \mu_U) - (\mu_{K-U+1} + \cdots + \mu_K)$$

is obtained by selecting the worst set of U arms, and the minimum non-zero sub-optimality gap $\Delta_{\min} \coloneqq \mu_U - \mu_{U+1}$ is obtained by replacing arm U in A_* with arm $U+1$, assuming that $\mu_U \neq \mu_{U+1}$.

Let $\nu = (p_1, \ldots, p_K)$ be an instance of `EgalMAB`. We say that ν is a 1-subgaussian `EgalMAB` if for all arms $a \in [K]$, $X \sim p_a$ is a 1-subgaussian random variable. Theorems 1 and 2 respectively provide problem-dependent and problem-independent upper bounds for the expected cumulative regret of running `EgalUCB` on a 1-subgaussian `EgalMAB`.

Theorem 1 (Problem-Dependent Upper Bound)**.** *Let (ν, T, U) a 1-subgaussian `EgalMAB`. After running `EgalUCB` for T time steps, we have*

$$R_T \leq \frac{2136(K-U)\ln(T)}{\Delta_{\min}} + \frac{4K\Delta_{\max}}{U}.$$

Theorem 2 (Problem-Independent Upper Bound)**.** *Let (ν, T, U) be a 1-subgaussian `EgalMAB` with $\mu_a \in [0,1]$ for all arms $a \in [K]$. After running `EgalUCB` for T time steps, we have*

$$R_T \leq \sqrt{\frac{8544(K-U)\,T\ln(T)}{U}} + \frac{4K\min\{U, K-U\}}{U}.$$

When the number of arms K and users U are fixed, the problem-independent upper bound increases with the number of time steps T at a rate of $O(\sqrt{T\ln(T)})$. Notably, when $U = 1$, the `EgalMAB` instance and the `EgalUCB` policy reduce to a classic MAB instance and the `UCB1` policy (Auer et al., 2002). As a sanity check, we achieve the bound of `UCB1` $O(\sqrt{T\ln(T)})$.

Next, we examine how the number of users U affects performance. Consider some fixed time horizon T and number of arms K. Since $T \gg K$, the first terms in both the problem-independent and problem-dependent upper bounds dominates the regrets. Furthermore, when $K = U$, every user would have played every arm exactly once after each block, yielding an expected cumulative reward of $b\mu_*$ after block b. By definition, this implies that the expected cumulative regret $R_T = 0$. This behavior is reflected in both upper bounds, since $K - U = 0$ and $\Delta_{\max} = 0$ when $K = U$.

Additionally, the problem-independent upper bound decreases as U approaches K, and this reduction scales with $O(1/\sqrt{U})$. There are two reasons for this. Firstly, if we fix some `EgalMAB` instance ν and vary U, then increasing U results in decreasing $T\mu_*/U$. Secondly, since `EgalUCB` assigns the arms in a round-robin fashion during each block, as long as the UCB values of the top arms are consistently among the highest regardless of their order, `EgalUCB` will also consistently select a good set of arms. This implies that as U increases, this problem becomes more statistically robust to the variability inherent in the estimates of the arms. Loosely speaking, the more users we have, the easier it is to match the performance of a policy always plays round-robin the set of arms A_*.

We further consider the scope for algorithmic improvement by deriving a policy-independent lower bound. Let $\nu = (p_1, \ldots, p_K)$ be an `EgalMAB` instance and π be any policy. We denote $R_{\pi\nu}$ as the expected cumulative regret of running π on ν for T time steps. Theorem 3 provides a policy-independent lower bound for the regret $R_{\pi\nu}$. This bound applies to the class $\mathcal{V}$ of all Gaussian `EgalMAB` instances $\nu = (p_1, \ldots, p_K)$ where, for all $a \in [K]$, the reward density $p_a = \mathcal{N}(\mu_a, 1)$ and $\mu_a \in [0,1]$.

Theorem 3 (Policy-Independent Lower Bound)**.** *Suppose $K \geq 2U$. For any policy π, there exist an* `EgalMAB` *instance $\nu \in \mathcal{V}$ with regret*

$$R_{\pi\nu} \geq \frac{\sqrt{(K-U)\,T}}{76U}.$$

The lower bound suggests that `EgalUCB` is tight in T up to logarithmic factors. This factor is expected when using a UCB-based policy due to the choice of confidence radius $\epsilon_{b,b'}$. Drawing parallels to the MOSS policy (Audibert & Bubeck, 2009) in classic K-armed MAB problems, we also conjecture that there exists a MOSS-based policy that can shave away the $\sqrt{\log T}$ term in the regret bound.

Furthermore, there is a multiplicative gap of $1/\sqrt{U}$ between the lower bound and the problem-independent upper bound. Our experimental results in Section 7 suggest that our upper bound analysis is not tight, as we empirically observe the $O(1/U)$ behavior when running `EgalUCB`.

6 Regret Analysis

In this section, we provide a proof sketch for the main results in Section 5. Detailed proofs for the upper bounds and lower bound can be found in Appendices A and B respectively.

6.1 Problem-Dependent Upper Bound

The regret upper bound in Theorem 1 consists of a sum of two terms: the first term is dependent on T and the second is independent of T. The first (resp. second) term arises from the regret accumulated over blocks where some *good* event $\mathcal{E}_b$ occurs (resp. did not occur). This *good* event $\mathcal{E}_b$ is the event that μ_a and its estimate $\hat{\mu}_a$ are at a distance of at most $\epsilon_{b-1,B_{a,b-1}}$ at the beginning of block b for all $a \in [K]$. Formally, we define

$$\mathcal{E}_b \coloneqq \bigcap_{a=1}^{K} \left\{ \left|\hat{\mu}_{a,B_{a,b-1}} - \mu_a\right| \leq \epsilon_{b-1,B_{a,b-1}} \right\}.$$

Note that we will often abuse the set notation when defining events. In particular, when we have a proposition P, we use $\mathcal{E} = \{P\}$ to signify that $\mathcal{E}$ is the set of all outcomes in the underlying probability space where P holds. Lemma 1 shows that $\mathcal{E}_b$ occurs with high probability.

Lemma 1. *Let (ν, T, U) be a 1-subgaussian* `EgalMAB`*. Then, for all blocks $b \in [B]$,*

$$\mathbb{P}(\mathcal{E}_b^{\mathsf{c}}) \leq \frac{2K}{b^2U^3}.$$

To facilitate analysis, we then consider another *good* event $\mathcal{F}_b$ for which the set of arms $A_b \subseteq_U [K]$ selected during block b is sub-optimal but "not too bad". This is defined as

$$\mathcal{F}_b \coloneqq \left\{ 0 < \Delta_{A_b} \leq 2 \sum_{a \in A_b^-} \epsilon_{B,B_{a,b-1}} \right\}.$$

Lemma 2 shows that if the set of arms A_b played during block b is sub-optimal and $\mathcal{E}_b$ occurs, then $\mathcal{F}_b$ must follow.

Lemma 2. *Let $b \in [B]$. If the set of arms A_b played during block b is sub-optimal and $\mathcal{E}_b$ occurs, then $\mathcal{F}_b$ also occurs.*

We proceed by partitioning the regret into two terms: one conditioned on the high-probability event $\mathcal{F}_b$ using Lemma 2 and another conditioned on the low-probability event $\mathcal{E}_b^{\mathsf{c}}$ using Lemma 1. This result is formally stated in Lemma 3.

Lemma 3. *Let (ν, T, U) be a 1-subgaussian* `EgalMAB`*. Then, after T time steps, for all users $u \in [U]$, we have*

$$R_{u,T} \leq \sum_{b=1}^{B} \mathbb{E}[\Delta_{A_b} \mathbb{I}\{\mathcal{F}_b\}] + \frac{\pi^2 K \Delta_{\max}}{3U^3}.$$

Let us now focus on bounding the contributions made by the high-probability term. We first partition $\mathcal{F}_b$ into countably many mutually exclusive events $\{\mathcal{G}_{b,i}\}_i$ so that we can decompose the sum

$$\sum_{b=1}^{B} \mathbb{E}[\Delta_{A_b} \mathbb{I}\{\mathcal{F}_b\}] = \sum_{b=1}^{B} \sum_{i=1}^{\infty} \mathbb{E}[\Delta_{A_b} \mathbb{I}\{\mathcal{G}_{b,i}\}].$$

To define the events $\{\mathcal{G}_{b,i}\}_i$, let $\alpha = 0.13$, $\beta = 0.22$, and

$$\gamma = 24 \left(\frac{1-\beta}{\sqrt{\alpha} - \beta} \right)^2$$

be constants that are carefully chosen, and define

$$m_{b,i} \coloneqq \frac{\gamma \alpha^i U \ln(BU)}{\Delta_{A_b}^2}$$

for all $i \in \mathbb{N}$. Let

$$L_{b,i} \coloneqq \{a \in A_b^- : B_{a,b-1} \leq m_{b,i}\}$$

to be the set of arms in A_b^- that are played for fewer than $m_{b,i}$ blocks at the beginning of block b. For convenience, let $L_{b,0} = A_b^-$. For each $b \in [B]$ and $i \in \mathbb{N}$, the event $\mathcal{G}_{b,i}$ is then defined as

$$\mathcal{G}_{b,i} \coloneqq \bigcap_{j=1}^{i-1} \big\{ |L_{b,j}| < \beta^j U \big\} \cap \big\{ |L_{b,i}| \geq \beta^i U \big\} \cap \big\{ \Delta_{A_b} > 0 \big\}.$$

It is clear that at most one of $\{\mathcal{G}_{b,i}\}_i$ can occur. However, to show that it is a partition for $\mathcal{F}_b$, we also need to show that at least one of $\{\mathcal{G}_{b,i}\}_i$ must happen. This is shown in Lemma 4.

Lemma 4. *Assume that $b > K/U$. On the event $\mathcal{F}_b$, exactly one of the events in $\{\mathcal{G}_{b,i}\}_i$ occurs.*

Using the newly-defined events $\{\mathcal{G}_{b,i}\}_{b,i}$, we bound the contributions of the the high-probability term in Lemma 5.

Lemma 5. *Let $\nu = (p_1, \ldots, p_K)$. Suppose that p_a is the density for a 1-subgaussian distribution for all $a \in [K]$. Then, after T time steps,*

$$\begin{aligned} \sum_{b=1}^{B} \mathbb{E}[\Delta_{A_b} \mathbb{I}\{\mathcal{F}_b\}] &\leq 2136 \ln(BU) \sum_{a \in \Lambda} \frac{1}{\Delta_{a,N_a}} + \frac{K \Delta_{\max}}{U} \\ &\leq \frac{2136(K-U)\ln(BU)}{\Delta_{\min}} + \frac{K\Delta_{\max}}{U}. \end{aligned}$$

for all users $u \in [U]$.

The proof of Theorem 1 simply involves substituting the result of 5 into Lemma 3.

Proof of Theorem 1. We can bound the regret by

$$\begin{aligned} R_{u,T} &\leq \sum_{b=1}^{B} \mathbb{E}[\Delta_{A_b} \mathbb{I}\{\mathcal{F}_b\}] + \frac{\pi^2 K \Delta_{\max}}{3U^3} \\ &\leq \frac{2136(K-U)\ln(BU)}{\Delta_{\min}} + \frac{K\Delta_{\max}}{U} + \frac{\pi^2 K \Delta_{\max}}{3U^3} \\ &\leq \frac{2136(K-U)\ln(T)}{\Delta_{\min}} + \frac{4K\Delta_{\max}}{U} \end{aligned}$$

where the first inequality holds due to Lemma 3 and the second inequality holds due to Lemma 5. □

6.2 Problem-Independent Upper Bound

Much like the expression in Theorem 1, the first term of the regret bound in Theorem 2 arises from the cumulative regret accumulated in blocks where the high-probability event $\mathcal{E}_b$ occurs. Conversely, the second term arises from the regret accumulated in the initial blocks and the blocks where $\mathcal{E}_b$ did not occur. The proof of Theorem 2 can be found in Appendix A.

6.3 Policy-Independent Lower Bound

Our proof for Theorem 3 consists of constructing two `EgalMAB` instances ν and ν' that are "close" enough that it is difficult for any policy to distinguish between them statistically, yet "far" enough that a sequence of actions that is good for one instance is bad for the other. Specifically, let

$$\Delta = \sqrt{\frac{K-U}{8TU^2}}$$

and set $\nu = (p_1, \ldots, p_K) \in \mathcal{V}$ to be an `EgalMAB` instance with

$$\mu_a = \begin{cases} \Delta, & \text{if } a \in [U] \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Assuming that $K \geq 2U$, let

$$A' = \operatorname*{arg\,min}_{A \subseteq_U [K] \setminus [U]} \sum_{a \in A} \mathbb{E}_{\pi\nu}[T_{a,T}]$$

be the set of U arms that are sub-optimal under ν have been played the fewest number of times under the distribution $\mathbb{P}_{\pi\nu}$. Set $\nu' \in \mathcal{V}$ to be an `EgalMAB` instance with

$$\mu'_a = \begin{cases} 2\Delta, & \text{if } a \in A' \\ \mu_a, & \text{otherwise.} \end{cases} \tag{2}$$

Let $\mathbf{\Gamma}$ be the set of size-U subsets of $[K]$ that contains at least $U/2$ arms from $[U]$, and let

$$\mathcal{H} \coloneqq \left\{ \sum_{A \in \mathbf{\Gamma}} T_{A,T} \leq \frac{T}{2} \right\}$$

be the event that, for at least $T/2$ time steps, the policy π selects a set in which at least half of it consists of arms from $[U]$. Lemma 6 uses the Bretagnolle–Huber inequality with $\mathcal{H}$ to reduce the lower bound computation to evaluating the KL-divergence between $\mathbb{P}_{\pi\nu}$ and $\mathbb{P}_{\pi\nu'}$.

Lemma 6. *Let ν and ν' be the `EgalMAB` instances defined by equation 1 and equation 2. Under the assumptions of Theorem 3, we have*

$$\begin{aligned} R_{\pi\nu} + R_{\pi\nu'} &> \frac{\Delta T}{4} \big(\mathbb{P}_{\pi\nu}(\mathcal{H}) + \mathbb{P}_{\pi\nu'}(\mathcal{H}^{\mathsf{c}}) \big) \\ &\geq \frac{\Delta T}{8} \exp\big(-D_{\mathrm{KL}}(\mathbb{P}_{\pi\nu} \| \mathbb{P}_{\pi\nu'}) \big). \end{aligned}$$

Part of the novelty in our analysis involves reducing the computation of this KL-divergence to a combinatorial problem. Lemma 7 decompose the KL-divergence from Lemma 6 into an expression that involves counting the number of times each arm $a \in A'$ is played, and Lemma 8 shows, using a counting argument, that the number of times an arm $a \in A'$ is played is at most $TU^2/(K-U)$.

Lemma 7. *Let ν and ν' be the `EgalMAB` instances defined by equation 1 and equation 2. Under the assumptions of Theorem 3, we have*

$$D_{\mathrm{KL}}(\mathbb{P}_{\pi\nu} \| \mathbb{P}_{\pi\nu'}) \leq 4\Delta^2 \sum_{a' \in A'} \mathbb{E}_{\pi\nu}[T_{a',T}].$$

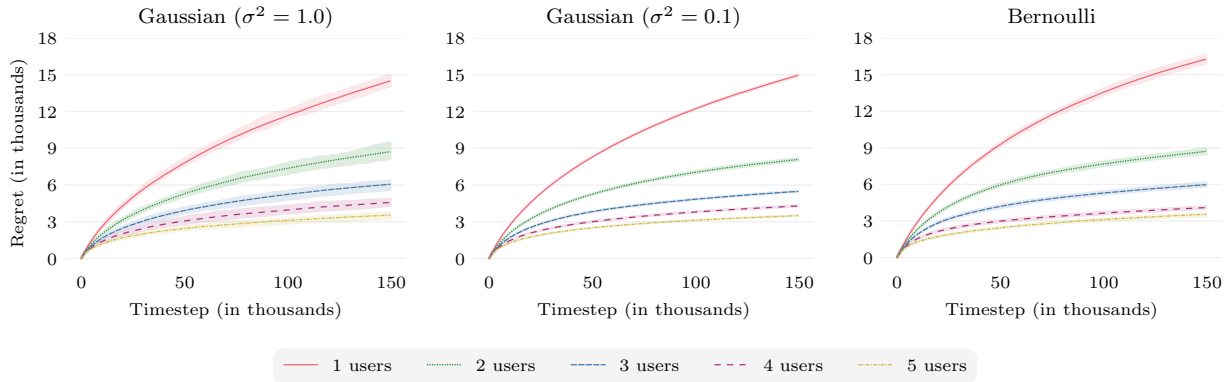

Figure 3: Expected regret incurred by `EgalUCB` over $T = 150{,}000$ time steps on simulated data with $K = 10$. Each line corresponds to a different U. The lighter region around each line represents the range between the minimum and maximum expected regrets observed over a total of 30 independent runs.

Lemma 8. *Let ν and ν' be the `EgalMAB` instances defined in equation 1 and equation 2. Under the assumptions of Theorem 3, we have*

$$\sum_{a \in A'} T_{a,T} \leq \frac{TU^2}{K - U}$$

almost surely.

The proof of Theorem 3 simply involves substituting the results of Lemma 7 and Lemma 8 into Lemma 6.

Proof of Theorem 3. We have

$$\begin{aligned} R_{T,\pi,\nu} + R_{T,\pi,\nu'} &\geq \frac{\Delta T}{8} \exp\left(-4\Delta^2 \sum_{a' \in A'} \mathbb{E}_{\pi\nu}[T_{a',T}]\right) \\ &\geq \frac{\Delta T}{8} \exp\left(-\frac{4\Delta^2 T U^2}{K - U}\right) \\ &= \frac{\Delta T}{8} \exp(-1/2) \\ &> \frac{\sqrt{T(K-U)}}{38U}. \end{aligned}$$

Since $2\max\{R_{T,\pi,\nu}, R_{T,\pi,\nu'}\} \geq R_{T,\pi,\nu} + R_{T,\pi,\nu'}$, dividing by 2 concludes the proof. □

7 Experiments

In this section, we present the results of our numerical experiments to validate the analysis of `EgalUCB`. These experiments include both a synthetic environment (Section 7.1) and real-world datasets (Sections 7.2 and 7.3). The code for these experiments can be found in the supplementary materials.

7.1 Synthetic Experiments

To empirically verify the problem-independent upper bound in Theorem 2, we conducted experiments on three synthetic datasets: Gaussian bandits with variance $\sigma^2 = 1.0$, Gaussian bandits with variance $\sigma^2 = 0.1$ and Bernoulli bandits.

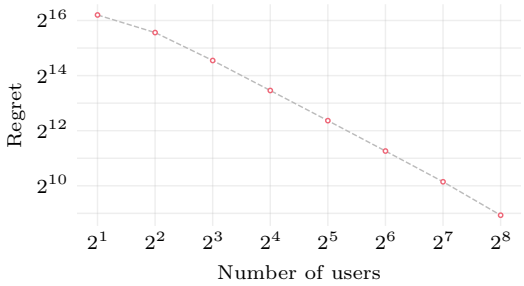

Figure 4: Expected regret incurred by `EgalUCB` over $T = 2^{18}$ time steps on Bernoulli bandits with $K = 2^{10}$ arms.

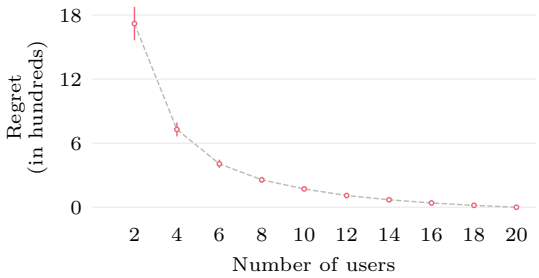

Figure 5: Expected regret incurred by `EgalUCB` over $T = 126{,}000$ time steps on Bernoulli bandits with $K = 20$ arms.

In each environment, we fixed $K = 10$ and varied U from 1 to 5. For each choice of U, we ran the experiment 30 times. In each run, we randomly generated the ground truth expected reward μ_a for each arm a using a uniform distribution with support $[0.01, 0.99]$. Figure 3 shows the expected regret incurred by `EgalUCB` over $T = 150{,}000$ time steps. As predicted by Theorem 2, the expected regret R_T is sub-linear in T and diminishes with U.

We conducted another experiment to verify the rate at which R_T diminishes with U when $K \gg U$. We fixed $K = 2^{10}$ and $T = 2^{18}$ while varying $U \in \{2^1, \ldots, 2^8\}$. We ran `EgalUCB` on instances of Bernoulli `EgalMAB` where

$$\mu_a = \begin{cases} 0.8 & \text{if } a \in [U] \\ 0.5 & \text{otherwise.} \end{cases}$$

This choice of μ_a ensures that μ_* is kept constant for all choices of U. Figure 4 shows the log-log plot of R_T against U. We observe that R_T diminishes with U at a rate of $O(U^{-c})$ for some constant $c \approx 1.0$. This corroborates with our policy-independent lower bound in Theorem 3 and suggests that our problem-independent upper bound in Theorem 2 may be loose.

To assess the rate at which R_T diminishes with U when $U \to K$, we ran a similar experiment with $T = 126{,}000$, $K = 20$, and $U \in \{2, 4, \ldots, 18, 20\}$ on Bernoulli `EgalMAB` instances. For each U, we ran the experiment 30 times. Figure 5 shows the plot of R_T against U. We observe that as U approaches K, the regret decreases to 0, and specifically when $U = K$, the regret is exactly 0. This observation aligns with the problem-independent upper bound in Theorem 2.

7.2 Google Cluster Usage Trace Dataset

The Google Cluster Usage Traces dataset comprises 2.4 TiB of compressed traces that record the workloads executed on Google compute cells (Wilkes, 2020). These traces are organized into tables that contain information about the machines and the instances running on them.

In our experiment, we focus on the *InstanceUsage* table from the `clusterdata_2019` trace. This table contains traces of both processor and memory usage during instance execution. To adapt to the `EgalMAB` setting, we designate each arm a as a machine, uniquely identifiable using the `machine_id` field in the table. We implicitly construct its reward distribution $\mathbb{P}_a$ by drawing an entry uniformly from the trace that corresponds to the machine and return the negative of the `cycles_per_instruction` field for its reward. Given the substantial size of the dataset, we will limit our analysis to the initial 4 million entries from the table. We pick $K = 100$ machines that contain the most amount of trace entries. Then, we formulate scenarios involving $U \in \{5, 10, 15, 20, 25\}$ *unseen* users. At each time step $t \in [T]$ where $T = 150{,}000$, we employ the `EgalUCB` policy to assign machines to these users.

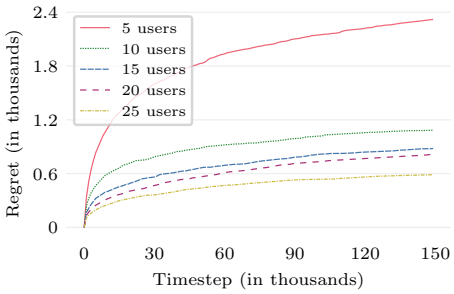

Figure 6: Expected regret incurred by `EgalUCB` over $T = 150{,}000$ time steps on the Google Cluster Usage Trace dataset with $K = 100$. Each line is associated with a different number of users U.

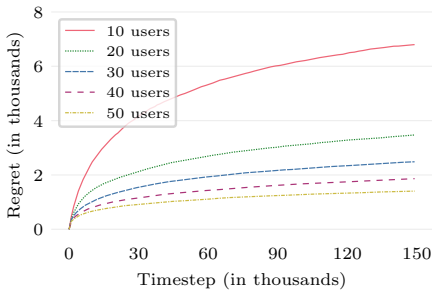

Figure 7: Expected regret incurred by `EgalUCB` over $T = 150{,}000$ time steps on the MovieLens 25M dataset with $K = 500$. Each line is associated with a different number of users U.

Figure 6 shows the expected regret R_T over time. Since the distribution $\mathbb{P}_a$ has bounded support, it is a subgaussian distribution. As such, this empirical finding dovetails with the $O(\sqrt{T \ln(T)})$ growth rate that Theorem 2 predicts.

7.3 MovieLens 25M Dataset

The MovieLens 25M dataset is widely used in recommender systems (Kużelewska, 2014; Forouzandeh et al., 2021) and collaborative filtering (He et al., 2017; Álvaro González et al., 2022) research. This dataset encompasses a substantial collection of 25,000,000 user ratings contributed by 162,000 users for a repository of 62,000 movies.

To adapt to the `EgalMAB` setting, we randomly select $K = 500$ movies and treat them as arms. For each movie $a \in [K]$, we implicitly construct its reward distribution $\mathbb{P}_a$ using the user ratings provided by existing users in the dataset. This empirical distribution is categorical and has support residing within $\{0.5, 1.0, \ldots, 4.5, 5.0\}$. Then, we formulate a scenario involving $U \in \{10, 20, 30, 40, 50\}$ *unseen* users. At each time step $t \in [T]$ where $T = 150,000$, we employ the `EgalUCB` policy to assign movies to these users.

Figure 7 shows the expected regret R_T over time. Similar to the Google Cluster Usage Trace dataset, the distribution $\mathbb{P}_a$ is subgaussian for all machines a. As such, this empirical finding aligns with the $O(\sqrt{T \ln(T)})$ growth rate that Theorem 2 predicts.

8 Conclusions

In this work, we introduced `EgalMAB`, an extension to the MAB framework with egalitarian consideration. The `EgalUCB` policy was proposed and shown to achieve an expected regret of

$$O\left(\sqrt{\frac{T \ln(T) \cdot (K - U)}{U}}\right).$$

We also derived a lower bound that matches the upper bound up to a multiplicative gap of $1/\sqrt{U}$ and a term logarithmic in T. Our experiments on simulated and real-world data validated the theoretical analysis. Our empirical results lead us to conjecture that `EgalUCB` is indeed tight with respect to U. This gap could potentially be reconciled with a more refined analysis of the upper bound in future work.

References

Álvaro González, Fernando Ortega, Diego Pérez-López, and Santiago Alonso. Bias and unfairness of collaborative filtering based recommender systems in movielens dataset. *IEEE Access*, 10:68429–68439, 2022. doi: 10.1109/ACCESS.2022.3186719.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32 (11):968–976, 1987. doi: 10.1109/TAC.1987.1104491.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, Berlin, Germany, 2009. Springer-Verlag.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. doi: 10.1023/A:1013689704352.

Orly Avner and Shie Mannor. Concurrent bandits and cognitive radio networks, 2014.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012. doi: https://doi.org/10.1016/j.jcss.2012.01.001.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J. Mach. Learn. Res.*, 17(1):1746–1778, Jan 2016.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. The fair contextual multi-armed bandit. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pp. 1810–1812, California, United States, 2020. International Foundation for Autonomous Agents and Multiagent Systems.

Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, pp. 299–308, Cambridge, United Kingdom, 2020. Association for Computing Machinery. doi: 10.1145/3319502.3374806. URL `https://doi.org/10.1145/3319502.3374806`.

Kaize Ding, Jundong Li, and Huan Liu. Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 357–365, New York, United States, 2019. Association for Computing Machinery. doi: 10.1145/3289600.3290964.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pp. 67–82, California, United States, 2018. PMLR. URL `https://proceedings.mlr.press/v85/durand18a.html`.

Saman Forouzandeh, Kamal Berahmand, and Mehrdad Rostami. Presentation of a recommender system with ensemble learning and graph embedding: A case on movielens. *Multimedia Tools and Applications*, 80(5):7805–7832, Feb 2021. doi: 10.1007/s11042-020-09949-5.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012. doi: 10.1109/TNET.2011.2181864.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 173–182, Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. doi: 10.1145/3038912.3052569.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24005–24017, New York, United States, 2021. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/c96ebeee051996333b6d70b2da6191b0-Paper.pdf`.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29, New York, United States, 2016. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf`.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 158–163, New York, United States, 2018. Association for Computing Machinery. doi: 10.1145/3278721.3278764.

Wassim Jouini, Damien Ernst, Christophe Moy, and Jacques Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*, Medenine, Tunisia, 2009. Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/ICSCS.2009.5412697.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, volume 37 of *ICML'15*, pp. 1152–1161, Lille, France, 2015. JMLR.org.

George Kousiouris, Tommaso Cucinotta, and Theodora Varvarigou. The effects of scheduling, workload type and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks. *Journal of Systems and Software*, 84(8):1270–1291, 2011. doi: https://doi.org/10.1016/j.jss.2011.04.013.

Urszula Kużelewska. Clustering algorithms in hybrid recommender system on movielens data. *Studies in Logic, Grammar and Rhetoric*, 37:125–139, Jan 2014. doi: 10.2478/slgr-2014-0021.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 535–543, California, United States, 2015a. PMLR. URL `https://proceedings.mlr.press/v38/kveton15.html`.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'15*, pp. 1450–1458, Massachusetts, United States, 2015b. MIT Press.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2020. doi: 10.1109/TNSE.2019.2954310.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010. doi: 10.1109/TSP.2010.2062509.

Jonas Mueller, Vasilis Syrgkanis, and Matt Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, New York, United States, 2019. Curran Associates Inc.

John Rawls. *A Theory of Justice: Original Edition*. Harvard University Press, Massachusetts, United States, 1971. URL `http://www.jstor.org/stable/j.ctvjf9z6v`.

Ayush Sawarni, Soumybrata Pal, and Siddharth Barman. Nash regret guarantees for linear bandits, 2023.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 974–980, California, United States, 2015. AAAI Press.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits, 2021.

Xuchuang Wang, Hong Xie, and John C. S. Lui. Multiple-play stochastic bandits with shareable finite-capacity arms. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23181–23212, Maryland, United States, 2022. PMLR. URL `https://proceedings.mlr.press/v162/wang22af.html`.

Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3026–3036, New York, USA, 2017. Curran Associates Inc.

John Wilkes. Yet more Google compute cluster trace data. Google research blog, Apr 2020. Posted at `https://ai.googleblog.com/2020/04/yet-more-google-compute-cluster-trace.html`.