Reinforcement Learning with Fairness Constraints for Resource Distribution in Human-Robot Teams

Houston Claure
Cornell University
United States
hbc35@cornell.edu

Yifang Chen
University of Southern California
United States
yifang@usc.edu

Jignesh Modi
University of Southern California
United States
jigneshm@usc.edu

Malte Jung
Cornell University
United States
mfj28@cornell.edu

Stefanos Nikolaidis
University of Southern California
United States
nikolaid@usc.edu

Abstract: Much work in robotics and operations research has focused on optimal resource distribution, where an agent dynamically decides how to sequentially distribute resources among different candidates. However, most work ignores the notion of fairness in candidate selection. In the case where a robot distributes resources to human team members, favoring heavily the highest performing teammate can have negative effects in team dynamics and system acceptance. We introduce a multi-armed bandit algorithm with fairness constraints, where a robot distributes resources to human teammates of different skill levels. In this problem, the robot does not know the skill level of each human teammate, but learns it by observing their performance over time. We define fairness as a constraint on the minimum rate that each human teammate is selected throughout the task. We provide theoretical guarantees on performance and perform a large-scale user study, where we adjust the level of fairness in our algorithm. Results show that fairness in resource distribution has a significant effect on users' trust in the system.

Keywords: Reinforcement Learning, Fairness, Multi-Armed Bandits, Trust

1 Introduction

We focus on the problem of resource distribution in human-robot teams. For instance, a factory robot assists several workers by delivering parts needed for an engine assembly. Some workers are experienced and fast, others are inexperienced and slow. The robot, unaware of each workers experience level, must decide how to distribute resources. If the robot is optimal in the traditional sense, it should first assign resources to learn the performance of each individual worker (exploration) and then assign as many pieces as possible to the most experienced worker (exploitation).

This approach, however, fails to account that assigning more resources to the highest performing worker may be perceived as unfair by the other worker. In turn, this may affect their perception of the interaction and trust in the system. In fact, previous work has shown that ignoring human preferences in task allocation can negatively affect users' willingness to work with the system [1]. Ultimately, team performance depends to a large degree on people's interpersonal orientation, i.e., how people perceive each other and interact with each other [2]. Groom and Nass [3] argue that our ability to build effective human-robot teams depends on a team's ability to build trust between all members of a team, and much work in human-robot interaction has focused on establishing perceived team fluency and trust in human-robot teams [4, 5, 6, 7, 8, 9].

We focus on the notion of *fairness* in resource distribution. We formalize how a robot can maximize performance, while guaranteeing that each human teammate will be assigned a minimum rate of resources at any given time throughout the task. Our thesis is that, by accounting for fairness in resource allocation, we can significantly improve users' trust in the system.

To this end, we cast the problem as a multi-armed bandit, where each human teammate is represented as an arm with an unknown reward function corresponding to their skill level. We then propose a *multi-armed bandit algorithm with fairness constraints*, which builds upon the standard Upper Confidence Bound (UCB) algorithm [10]. We propose a stochastic version of the algorithm, where a minimum pulling rate for each arm is satisfied in expectation, and a deterministic version where the constraint is strictly satisfied anytime throughout the task. We provide theoretical guarantees of performance in the form of regret bounds for both algorithms.

To assess the effect of fairness on users' perception of the interaction, we execute a large-scale user study on a Tetris game, where two players are sequentially assigned a batch of blocks by the algorithm. We implement the algorithm with three levels of fairness, representing the required minimum allocation rate for each player: 25%, 33% and 50%.

Results show that fairness affects significantly the trust of the players that performed worse than their teammates: those in the 33% condition trusted the system significantly more, compared to the 25% condition. Surprisingly, we did not observe a decrease in performance in the fairer distributions, even though the stronger player was selected less frequently. On the contrary, the average scores were higher when fairness increased. These results improve our understanding of the theory and implications of fairness in resource distribution in human-robot teams.

2 Background

2.1 Stochastic Multi-Armed Bandits

The stochastic multi-armed bandits (MAB) framework without a minimum pulling rate requirement has been theoretically well studied. The gambler is tasked with choosing an arm, i, from K arms at each time step $t = 1, 2, 3, ., n$. At every time t, the gambler pulls an arm $i_t \in [K]$ while simultaneously the environment decides the reward vector $r_t \in [0, 1]^K$ from a fixed distribution with expectation $\mathbb{E}[r_t] = \mu$. The gambler, however, can only observe $r_t(i_t)$ but not the whole vector. Therefore, the gambler's goal is to pull the sequence of arms, based on the past information, that can maximize the overall accumulated reward.

The best arm in hindsight is defined as $i^* = \text{argmax}_{i \in [K]} \mu(i)$ and $\mu^* = \mu(i^*)$. We use regret to measure the performance of this algorithm, which is how worse our algorithm performs compared to the benchmark strategy – always pulling the best arm in each step.

$$Reg_T = T\mu^* - \sum_{t=1}^{T} \mu(i_t)$$

An optimal solution to such a problem was proposed as the Upper Confidence Bound (UCB). It was originally introduced by Lai and Robbins [11] and expanded by Agrawal [12]. Building upon these works, Auer, Cesa-Bianchi & Fisher [10] introduced the Upper Confidence Bound Algorightm (UCB). At its most basic form of this algorithm, at each time t, we want to estimate the expected reward of each arm by using the mean of its empirical rewards in the past and the number of times it has been pulled, which gives us a confidence interval that the arm will lie in. Then the algorithm proceeds to pick the arm with largest estimated expected reward.

This work has inspired a family of upper confidence bound variant algorithms for an array of different applications [13, 14, 15, 16, 17]. For a review of these algorithms we point readers to [18].

More recent work regarding multi-armed bandits has seen applications towards the improvement of human-robot interaction. Recent work by Chan et al. has investigated using a MAB algorithm for the use of an assistive robotic system with the goal of exploring human preferences [19] and assisting human learning [20].

Of particular relevance is very recent work on sleeping bandits with fairness constraints [21], in a setting where multiple arms can be played simultaneously and some arms may be unavailable. Fairness is defined as a minimum rate satisfied in expectation and at the end of the task, whereas in our work we require the rate to be satisfied strictly and anytime throughout the task. Fairness in the context of MABs has also been studied in Joseph et al. [22]. The definition of fairness there is quite different, in that a worse arm should not be picked compared to a better arm, despite the uncertainty

on payoffs. Their proposed algorithm chooses two arms with equal probability, until it has enough data to deduce the best of the two arms.

2.2 Fairness in Resource Distribution

Human collaboration has been shown to have strong links to fairness [23]. While the actual definition of fairness is debated, we aim to look at it within the scope of resource distribution. In this context, fairness relates to the degree of allocation that a resource is given to an individual within a group [24]. While an equal distribution of resources across all members within a group seems ideal, researchers [25, 26] have shown that inequalities are deemed appropriate, particularly when they optimize the outcome of the group. On the other hand, perceived inequalities have a strong impact on individuals' behavior, often motivating them to act contrary to their rational self-interest with the goal of eliminating the inequality [27, 28]. Previous work has shown that perceived fairness affects job satisfaction [29] and can induce retaliation behavior from the affected party [30]. Interestingly, individuals perceive fairness differently when decisions are made by an algorithm, compared to a human [31, 32].

3 Algorithm

The original unconstrained UCB algorithm fails in ensuring "fairness" because when time passes, a large set of bad arms will hardly be used again. Therefore, we propose two new algorithms with optimal regret bound guarantees. Both of them are based on the unconstrained UCB algorithm, where we adopt the idea of estimating the expected reward of each arm by using the mean of its empirical rewards in the past and the number of times it has been pulled. We prove all theorems in the Appendix.

3.1 Strict-rate-constrained UCB Algorithm

Definition 1 *Let S be any K-elements set that all its elements are drawn from $[\frac{1}{v}]$ without replacement. Then define $g : S \to [K]$ as some one-to-one function.*

Algorithm 1: Strictly-rate-constrained UCB

Input: time horizontal T, arm set $[K]$, minimum pull rate v

Definition: Denote $UCB_t(i) = \frac{1}{t-1}\sum_{s=1}^{t-1} r_t(i)\mathbf{1}\{i_s = i\} + 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}$, and τ_j be the beginning time of block j.

Initialize: $t = 1, j = 1, \tau_1 = K + 1, \tau_j = \tau_1 + \frac{j-1}{v}$.

while $t \leq K$ **do**

  Pull arm $i_t = t$

  $t \leftarrow t + 1$.

for $j = 1, 2, 3, \ldots$ **do** ▷ j indexes a block

  while $t < \tau_{j+1}$ **do**

    If $t - \tau_j + 1 \in S$, then pull the arm $i_t = g(t - \tau_j + 1)$,

    Otherwise, pull the arm $i_t = \mathrm{argmax}_{i\in[K]} UCB_t(i)$

    $t \leftarrow t + 1$

This algorithm guarantees that IN PRACTICE the pulling rate at any time for each arm is at least $v-\epsilon$, by fixing certain time slots where the algorithm will pull the prescheduling arms. Here $\epsilon = 1/t$. In other time slots, the algorithm will behave just like the normal UCB.

To be specific, in our algorithm, we divide T into blocks with length $\frac{1}{v}$. Our algorithm is flexible in that there are multiple choices of S and g that satisfy the minimum rate constraint. For example in the user study in section 4, when $v = \frac{1}{4}$ and $K = 2$, we choose $S = \{1, 3\}$ and $g(1) = 1, g(3) = 2$, which means we always pull arm 1 at τ_j and arm 2 at $\tau_j + 2$ for all j.

Now the benchmark strategy for pulling arm is always pulling the best arm in those non-prescheduled time slots. So the regret definition in this case becomes:

$$Reg = \mathbb{E}_{env}\left[\sum_{t\in\mathcal{I}} r(i^*) - r_t(i_t)\right]$$

where $\mathcal{I}$ is all the non-prescheduled time slots among T.

Theorem 1 *By running Alg. 1, we obtain the regret bound that is very close to the original unconstrained UCB,*

$$Reg_T \leq \sum_{i:\Delta_i>0}\left[\frac{16\ln T}{\Delta_i}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2(1-Kv)^2\Delta_i\right] + \mathcal{O}(K)$$

If $\Delta_i \in [0,1]\ \forall i$, we also get the worst case guarantee,

$$Reg_T \leq \mathcal{O}(\sqrt{TK\ln T} + K\ln T)$$

3.2 Stochastic-rate-constrained UCB Algorithm

Algorithm 2: Stochastic-rate-constrained UCB

Input: time horizontal T, arm set $[K]$, minimum pull rate v

Definition: Denote $UCB_t(i) = \frac{1}{t-1}\sum_{s=1}^{t-1} r_t(i)\mathbf{1}\{i_s = i\} + 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}$.

Initialize: $t = 1, j = 1, \tau_1 = K+1, \tau_j = \tau_1 + \frac{j-1}{v}$.

while $t \leq K$ **do**

  Pull arm $i_t = t$

  $t \leftarrow t+1$.

for $t = K+1, K+2, K+3, \ldots$ **do**

  With probability $1-Kv$, pull the arm $i_t = \text{argmax}_{i\in[K]}\, UCB_t(i)$,

  Otherwise, uniformly pull an arm i_t from all K arms

This algorithm guarantees that the EXPECTED pulling rate at any time for each arm is at least v. Instead of rescheduling some arms as in the deterministic algorithm above, this algorithm introduces some randomness. At each time t, we ensure that each arm has at least v probability to be pulled; while with $1-Kv$ probability, the algorithm will again pull the arm with the best estimated expected reward. We denote this distribution over arms as p_t where $p_t(\text{argmax}_{i\in[K]}\, UCB_t(i)) = (1-Kv)+v$ and $p_t(i) = v, \forall i \in [K] \setminus \text{argmax}_{i\in[K]}\, UCB_t(i)$.

In this case, the benchmark strategy is pulling the best estimated arm with probability $(1-Kv)$ at time t otherwise uniformly drawing a random arm. We present this strategy with the distribution p^* over K arms where $p^*(i^*) = (1-Kv)+v$ and $p^*(i) = v, \forall i \in [K] \setminus i^*$. So the regret definition in this case becomes:

$$\mathbb{E}_{env,learner}\left[\sum_{t=1}^{T} \mathbb{E}_{i_t\sim p^*}[r_t(i_t)] - \mathbb{E}_{a_t\sim p_t}[r_t(i_t)]\right]$$

Theorem 2 *By running Alg. 2, we obtain the regret bound that is very close to the original unconstrained UCB,*

$$Reg_T \leq \sum_{a:\Delta_i>0}\left[\min\left\{\frac{16\ln T}{\Delta_i} + (1-Kv)\Delta_i, (1-Kv)\Delta_i T\right\}\right]$$

If $\Delta_i \in [0,1]\ \forall i$, we also get the worst case guarantee,

$$Reg_T < \mathcal{O}(\sqrt{TK\ln T} + K\ln(T))$$

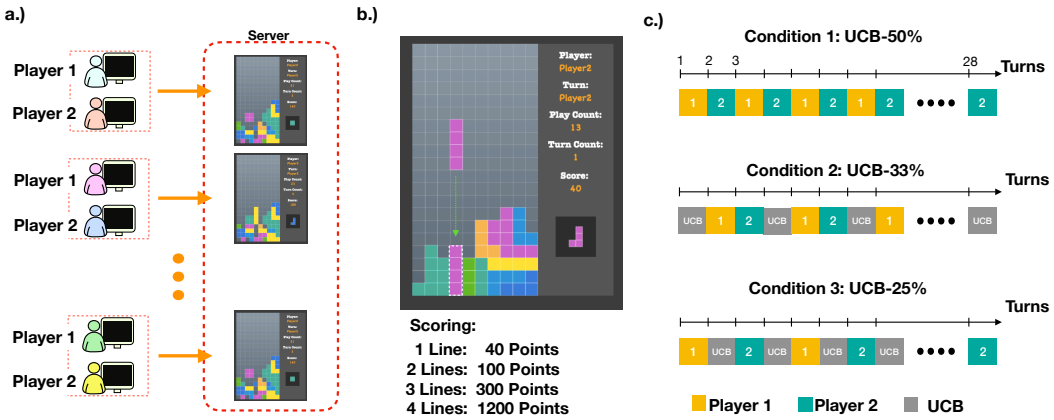

Figure 1: (a) Pairs of two remote human participants were connected to our cooperative Tetris game online. (b) The Tetris game followed the standard rules with two slight modifications. A different scoring metric as shown and the fact that only one participant had access to control the pieces per turn. (c) A visual representation of the three separate patterns that each condition offered.

4 User Study

We wish to assess the effect of implementing the algorithm with different fairness constraints in a setting where resources are distributed to human team members. We focus on a setting with two human teammates taking turns in completing a task. The system chooses at each turn which teammate should complete the task. We characterize the player that has the best performance of the two, as observed at the end of the task, as *strong* and the the other player as *weak.* The challenge of balancing between choosing the historically best player or a sub-optimal player allows us to investigate the impact of the system's decision on team performance, perceived fairness and trust in the system.

We make the following hypotheses.

H1: *Fairness will have a significant effect on perceived fairness and trust in the system of the weak players.* We focus on the weak players, since previous work [30] has shown that in unfair situations, adverse reactions mainly occur by the affected party.

H2: *Fairness will have a significant effect on team performance.* The fairer distributions favor the weak players, since they impose a constraint on the minimum number of pulls for both players. We expect that this will result in worse performance, compared to the less fair distributions that favor the strong player of the team.

4.1 Methodology

We developed a collaborative Tetris game and paired teams of two humans with a computer system running our MAB algorithm. The Tetris game challenges the spatial reasoning, reflex and decision speed of each individual. Each team's objective is to clear as many filled rows as possible by manipulating falling geometric pieces with the goal of obtaining the largest score under the allotted time frame. Our study had three conditions, each corresponding to a different fairness level, represented as a constraint on the minimum rate of arm pulls, that is on the minimum rate that a player is selected to play: 50%-UCB, 33%-UCB and 25%-UCB. We used a between-subjects design to avoid ordering effects. The system design is shown in Fig. 1.

We formally define our scenario as follows. The number of players in each game is set as $P = \{1, 2\}$ over a time horizon of $T = \{1, 2, ..., 30\}$. At each time step $t \in T$, seven consecutive Tetris pieces are allotted to a player $p \in P$. In the turns where the UCB algorithm was run, the upper confidence

Factor	Question Number	Question
Decision Fairness	Q1	How fair or unfair is it for **your partner** that the computer gave them the designated number of Tetris pieces?
	Q2	How fair or unfair is it for **you** that the robot gave you the designated number of tetris pieces?
Trust	Q1	How much do you trust the computer to make a good quality decision in the distribution of the tetris pieces?

Table 1: The subjective metrics that were used in our user study.

bound was calculated using as reward $r_{p,t}$:

$$r_{p,t} = \frac{S_{p,t}}{M * n_{p,t}}$$

where S_p is the score achieved by player p up to turn t, $n_{p,t}$ is the number of plays and M is a maximum value that we selected for normalization. After multiple pilot sessions we empirically set it to 300.

Measures: Information regarding an individual's performance was stored in a database during game play. We collected each player's individual score as well as the number of turns that was allocated to them. Additionally, we obtained the total score that each team accumulated at the end of the game play. Fig 1(b) shows the scoring convention that we used.

To measure levels of perceived decision fairness and trust we adapted survey questions from [33] (Table 1). Each response was measured on a seven-point Likert scale.

Procedures: We recruited participants using Amazon Mechanical Turk and utilized Qualtrics to create and collect survey question responses. Upon entering basic demographic information, AMT participants were instructed that they would be paired with a human partner and a computer system that would decide who has control of the falling pieces. It was further stated that their objective was to obtain the largest possible team score in the allotted turns. Our developed game platform then selected pairs of remote players to begin the game. Following standard Tetris rules, a player could rotate, speed up, or drop each falling pieces during a time step. We defined each time step as a set of seven consecutive falling pieces of which only the selected player could control. At the end of the time step the 50%, 33%, or 25% UCB, depending on condition, algorithm would run to select the next player. To reduce variance from sampling, we used the strict-rate-constrained UCB Algorithm (Alg. 1).

Fig. 1(c) shows the pattern of the distribution that was seen across each condition. This pattern was repeated for 30 time steps, with the exception of the first two time steps where each player played once. Each team was exposed to 210 pieces total. A code was given to participants upon the completion of the 30 rounds which enabled them to continue the Qualtrics survey.

Participants: We recruited a total of 290 participants from AMT and paid $1.00 for their participation in the task. 8 data points were removed as they failed our matching the AMT unique ID with the one given on Qualtrics. The final dataset contained N = 94, 98, 90 participants for UCB-50%, UCB-33%, and UCB-25% respectively (156 female, 124 male, 1 other, 1 did not disclose). Participants were recruited if they could speak English, were from the United States, and had previous ratings of 95% or higher. The average age of participants was 36 years old (SD = 11).

4.2 Analysis

Subjective: We grouped the subjective responses of each pair of players based on their comparative performance in the game (Figure 2). We focus the analysis on the weak players, that is the players that performed worse than their teammate. We present the responses of the strong players as well for completeness.

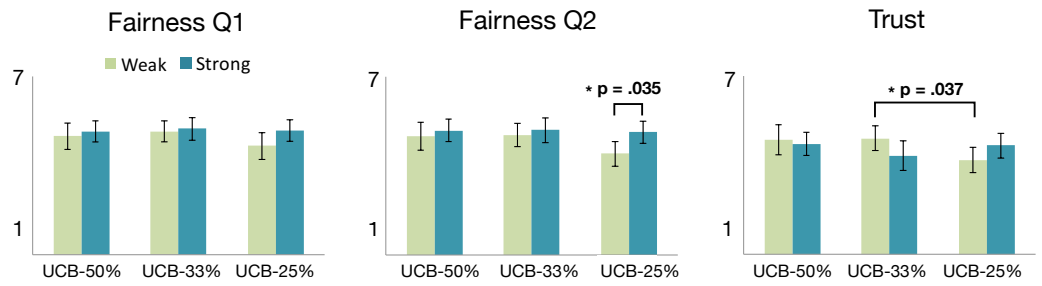

Figure 2: Responses to the subjective questions, grouped by player performance across each condition. Error bars represent the 95% confidence intervals.

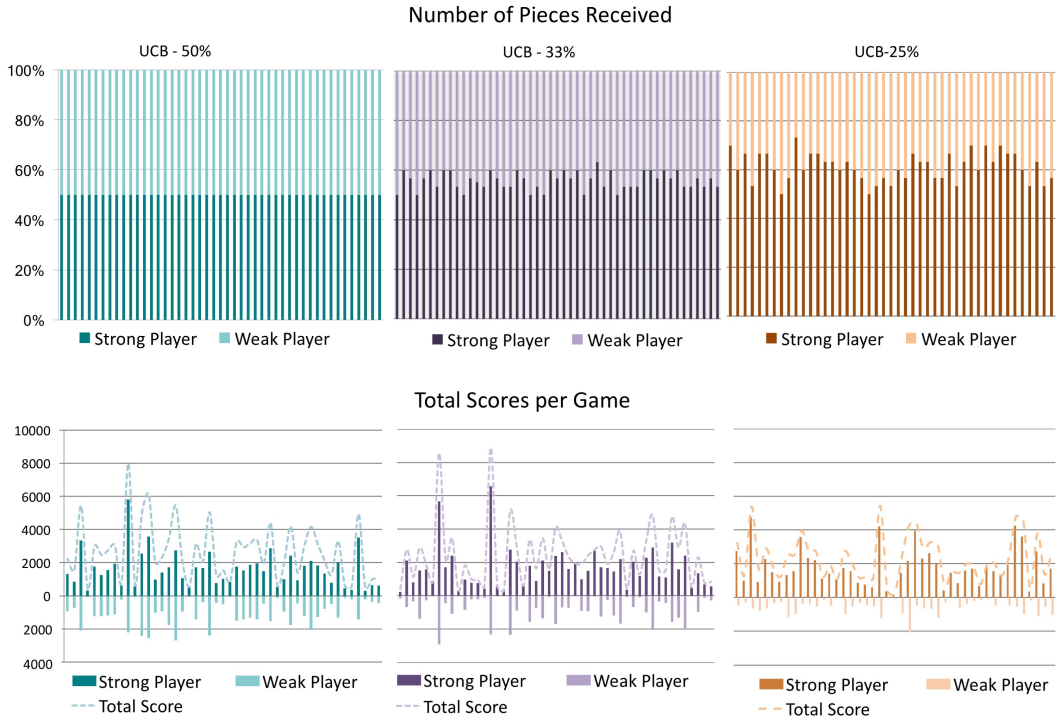

Figure 3: (Top) Percentage of the number of pieces that each player received. Each bar represents a separate game. (Bottom) Box plots of the total scores that each player individually and both players together achieved.

A one-way ANOVA was performed for weak players across all conditions (UCB-50% vs. UCB-33% vs. UCB-25%) for each subjective metric. Analysis indicates a significant effect of the reported trust score of the weak players across the three conditions ($F(2, 138) = 3.172, p = 0.025$). A Tukey HSD with adjusted p-values demonstrated higher trust ($p = 0.037$) towards the system running the UCB-33% compared to the UCB-25%. While trust scores in UCB-50% were higher than in the UCB-33%, the difference was not significant ($p = 0.081$). We found no other significant results in the other factors.

Post-hoc Analysis. We observe in Fig. 2 a noticeable difference in the responses between the strong and the weak players for different fairness conditions. Therefore, we conducted a post-hoc experimental analysis of the data to assess whether the responses between the weak and strong players varied significantly depending on the executed algorithm. Indeed, a 2 x 3 ANOVA with strength (weak vs. strong) and rate (UCB-50% vs. UCB-33% vs. UCB-25%) showed a main effect of play-

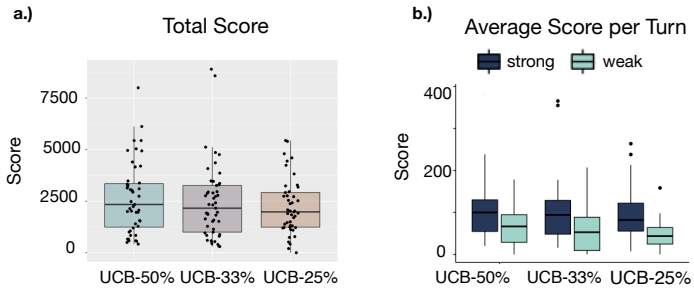

Figure 4: (a) Total scores for each condition. (b) Average score per turn for each condition.

ers' strength for Decision Fairness Q2 ($F(1, 276) = 4.778, p = 0.0297$). There were no interaction effects. Post-hoc comparison with Bonferroni corrections looking at strength within the different fairness levels, showed that weak players (M = 3.97, σ= 1.68) reported significantly lower ratings on fairness (Q2) than their strong counterparts (M = 4.82, σ = 1.49) in the UCB 25% condition ($p = 0.035$), which was the least fair condition. We observed no significant difference in perceived fairness between strong and weak players in the other two conditions.

These results show that in the least fair condition, there was a significant difference between the weak and the strong players in their perception of fairness. We also observe that reducing the minimum rate from 33% to 25% had a negative effect on the trust of weak players. On the other hand, Fig. 2 shows that trust scores between the UCB-50% and UCB-33% conditions were comparable.

To interpret these results, we observe the number of pieces received (arm pulls) for each condition in Fig. 3. In the UCB-50% condition, all players received equal number of pieces regardless of their performance. In the UCB-33% condition, while the strong players received more pieces, the difference with the weak players was small. On the other hand, in the UCB-25% condition there were several games where the weak players received less than 30% of the pieces, resulting in lower reported trust in that condition.

Objective: A one-way ANOVA on the performance of the two-player teams across the three conditions indicated no statistical significance. In fact, Fig. 4(a) shows that the medians of the total scores were higher for increasing levels of fairness. Plotting the individual scores of the players for each game in Fig. 3(bottom) illustrates this tendency as well.

This result does not match our initial hypothesis. To interpret this result, we plot the average scores per turn for each condition in Fig. 4(b). The average scores indicate how well the players performed on average every time they took a turn. Interestingly, we see that the distribution of the weak players' scores shifts towards lower scores as fairness decreases. While this result warrants further investigation, it indicates that assigning significantly less pieces to one of the players may negatively affect their performance, in addition to reducing their trust in the system. It showcases the importance of fairness when making resource distribution decisions.

5 Discussion

Limitations. Our work is limited in many ways. Algorithm 1 allows for a set of possible schedules that satisfy the minimum rate constraint, based on our choice of S and g. For instance, for UCB-25%, we chose to play the arm with the highest UCB bound in the second and fourth timeslot, but we could also select the first and second timeslot. In fact, given a minimum rate v there are $\frac{(1/v)!}{(1/v-K)!}$ permutations, and we have not captured the effect of different schedules within a fairness condition. Our model also does not capture changes in the performance of the individual players over time and it is worth exploring MAB algorithms that do not assume stationary rewards [34].

Implications. Fairness in resource distribution will play an important role in human-robot team dynamics and we are excited to have brought about a better understanding of the relationships between fairness, performance and trust in the system. We are also excited in exploring applications of these ideas to manufacturing and assistive care settings, where a robot distributes resources to multiple users.

References

[1] M. C. Gombolay, C. Huang, and J. A. Shah. Coordination of Human-Robot Teaming with Human Task Preferences. *AAAI Fall Symposium Series on AI-HRI*, 2015.

[2] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

[3] V. Groom and C. Nass. Can robots be teammates? Benchmarks in human-robot teams Identifying the best model for human-robot interaction. Technical report, 2007. URL http://www.cs.cmu.edu/~illah/CLASSDOCS/groom2007.pdf.

[4] M. C. Gombolay, R. A. Gutierrez, S. G. Clarke, G. F. Sturla, and J. A. Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39(3):293–312, 2015.

[5] J. Shah, J. Wiken, B. Williams, and C. Breazeal. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, 2011. ISBN 9781450305617. doi:10.1145/1957656.1957668.

[6] C.-M. Huang, M. Cakmak, and B. Mutlu. Adaptive Coordination Strategies for Human-Robot Handovers. Technical report. URL http://www.roboticsproceedings.org/rss11/p31.pdf.

[7] J. Baraglia, M. Cakmak, Y. Nagai, R. Rao, and M. Asada. Initiative in robot assistance during collaborative task execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 67–74. IEEE, 3 2016. ISBN 978-1-4673-8370-7. doi:10.1109/HRI.2016.7451735. URL http://ieeexplore.ieee.org/document/7451735/.

[8] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 307–315. ACM, 2018.

[9] H. Soh, P. Shu, M. Chen, and D. Hsu. The transfer of human trust in robot capabilities across tasks. *arXiv preprint arXiv:1807.01866*, 2018.

[10] P. Auer and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem*. Technical report, 2002. URL https://link.springer.com/content/pdf/10.1023/A:1013689704352.pdf.

[11] T. L. Lai Andherbertrobbins. Asymptotically Efficient Adaptive Allocation Rules*. Technical report, 1985. URL https://core.ac.uk/download/pdf/82425825.pdf.

[12] R. Agrawal. Sample mean based index policies by ¡i¿O¡/i¿ (log ¡i¿n¡/i¿) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 12 1995. ISSN 0001-8678. doi:10.2307/1427934. URL https://www.cambridge.org/core/product/identifier/S0001867800047790/type/journal_article.

[13] O.-A. Maillard, R. Munos, and G. Stoltz. A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. Technical report, 2011. URL http://proceedings.mlr.press/v19/maillard11a/maillard11a.pdf.

[14] R. Kleinberg, A. Slivkins, and E. Upfal. *Multi-Armed Bandits in Metric Spaces*. ISBN 978-1-60558-047-0. URL https://www.cs.cornell.edu/~rdk/papers/bandits-lip.pdf.

[15] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. Technical report, 2012. URL https://arxiv.org/pdf/1003.0146.pdf.

[16] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient Optimal Learning for Contextual Bandits. Technical report. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2011/01/DudikEtAl11full.pdf.

[17] A. Garivier. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems Eric Moulines. Technical report, 2008. URL https://arxiv.org/pdf/0805.3415.pdf.

[18] G. Burtini, J. Loeppky, and R. Lawrence. A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit. Technical report, 2015. URL https://arxiv.org/pdf/1510.00757.pdf.

[19] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan. The Assistive Multi-Armed Bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 354–363. IEEE, 3 2019. ISBN 978-1-5386-8555-6. doi:10.1109/HRI.2019.8673234. URL https://ieeexplore.ieee.org/document/8673234/.

[20] R. Pandya, S. H. Huang, D. Hadfield-Menell, and A. D. Dragan. Human-AI Learning Performance in Multi-Armed Bandits. Technical report. URL www.aaai.org.

[21] F. Li, J. Liu, and B. Ji. Combinatorial sleeping bandits with fairness constraints. *arXiv preprint arXiv:1901.04891*, 2019.

[22] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[23] J. R. Hackman and G. R. Oldham. Motivation through the design of work: Test of a theory. *Organizational behavior and human performance*, 16(2):250–279, 1976.

[24] T. Lan, D. Kao, M. Chiang, and A. Sabharwal. *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

[25] P. A. M. V. Lange, D. Batson, E. De Bruin, S. Koole, and A. M. V. Lange. The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation. Technical Report 2, 1999. URL https://pdfs.semanticscholar.org/8065/a450a90cff1d4464d156906924ec3dfd03b1.pdf.

[26] R. Fisman, S. Kariv, and D. Markovits. We study individual preferences for giving. Our experiments employ a graphical interface that allows subjects to see geometric representa-tions of choice sets on a computer screen and to make decisions through a simple point-and-Individual Preferences for Giving. Technical report. URL https://sites.bu.edu/fisman/files/2015/11/AER07-IPG.pdf.

[27] J. S. Lecture, E. Fehr, and A. Falk. Psychological foundations of incentives. Technical report, 2002. URL www.elsevier.com/locate/econbase.

[28] C. Camerer. *Behavioral game theory : experiments in strategic interaction*. Russell Sage Foundation, 2003. ISBN 9780691090399. URL https://press.princeton.edu/titles/7517.html.

[29] D. B. Mcfarlin and P. D. Sweeney. Distributive and Procedural Justice as Predictors of Satisfaction with Personal and Organizational Outcomes. Technical Report 3, 1992. URL https://www.jstor.org/stable/256489.

[30] D. P. Skarlicki and R. Folger. Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of applied Psychology*, 82(3):434, 1997.

[31] M. K. Lee and S. Baykal. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. doi:10.1145/2998181.2998230. URL http://dx.doi.org/10.1145/2998181.2998230.

[32] M. K. Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

[33] M. K. Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):205395171875668, 6 2018. ISSN 2053-9517. doi:10.1177/2053951718756684. URL http://journals.sagepub.com/doi/10.1177/2053951718756684.

[34] O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

6 Appendix

6.1 Notations

Some notations have already been defined in the main section, but for clarity, we will restate here: Denote $\Delta_i = \mu_{i^*} - \mu_i$. Let $\hat{\mu}_t(i) = \frac{1}{t-1}\sum_{s=1}^{t-1} r_t(i)$ be the mean of empirical rewards for arm i at time t so far and let $n_{t-1}(i)$ be the total number of times arm i has been pulled before time t. So $UCB_t(i) = \hat{\mu}_t(i) + \sqrt{\frac{\ln T}{n_{t-1}(i)}}$ and $i_t = \text{argmax}_{i\in[K]} UCB_t(i)$.

6.2 Auxiliary Theorems and Lemmas

Theorem 3 (Hoeffding's Inequality) *Let $X_1, ..., X_T \in [-B, B]$ for some $B > 0$ be independent random variables such that $\mathbb{E}[X_t] = 0, \forall t \in [T]$, then we have for all $\delta \in (0, 1)$,*

$$Pr(\frac{1}{T}\sum_{t=1}^{T} X_t \geq B\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}) \leq \delta$$

Lemma 1 *For all arm i, if the possible value range of $n_{t-1}(i)$ is $[k_s, k_e]$, then*

$$Prob\left[\mu(i) - \hat{\mu}_t(i) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}\right] \leq \sum_{k=k_s}^{k_e} \frac{1}{T^2}$$

Proof: We want to bound them by Hoeffding's Inequality, however, one trap here is that $n_{t-1}(i)$ is actually a random variable depending on all the rewards decided by the environment. To deal with this issue, imagine there is a infinite sequence of $X_1(i), X_2(i)...$ of independent samples of $\mathcal{D}_i$ for each action i and at time t observed reward $r_t(i_t)$ is the $n_t(i_t)$-th sample of this sequence, that is, $r_t(i_t) = X_{n_t(i_t)}(i_t)$. So $\hat{\mu}_{t-1}(i)$ as be written as $\tilde{\mu}_{n_{t-1}(i)}(i) = \frac{1}{n_{t-1}(i)}\sum_{k=1}^{n_{t-1}(i)} X_k(i)$.

So now we want to know what is the possible value of $n_{t-1}(i)$. According to the assumption $n_{t-1}(i) \in [k_s, k_e]$, we have,

$$\begin{aligned}
&Prob\left[\mu(i) - \hat{\mu}_{t-1}(i) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}\right] \\
&\leq Prob\left[\exists k \in [k_s, k_e] \quad s.t. \mu(i) - \tilde{\mu}_k(i) \geq 2\sqrt{\frac{\ln T}{k}}\right] \\
&\leq \sum_{k=k_0}^{k_e} Prob\left[\mu(i) - \tilde{\mu}_k(i) \geq 2\sqrt{\frac{\ln T}{k}}\right] \\
&\leq \sum_{k=k_0}^{k_e} \frac{1}{T^2}
\end{aligned}$$

The penultimate inequality is by hoeffding's inequality. □

6.3 Proof for Algorithm 1

6.3.1 Notations

We define $\mathcal{I}$ as the set of "non-prescheduled" time slots among T. Let $m_{t-1}(i)$ be the total number of times arm i has been pulled before time t and among $\mathcal{I}$, so $m_{t-1}(i) \leq n_{t-1}(i)$. Also $\mathcal{I}[i]$ means the j-th time slot in $\mathcal{I}$.

6.3.2 Main Proof

First we rewrite this regret in the form of variable Δ_i and m_T,

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}} r_t(i^*) - r_t(i_t)\right] = \mathbb{E}\left[\sum_{t\in\mathcal{I}} \mu^* - \mu(i_t)\right]$$

$$= \mathbb{E}\left[\sum_{t\in\mathcal{I}} \Delta_{i_t}\right]$$
$$= \sum_{i\neq i^*} \Delta_i \mathbb{E}\left[m_T(i)\right]$$

Here the first expectation is regarding to the whole environment randomness through T. The first equality comes from $\mathbb{E}_{\text{env at t}}(r_t) = \mu$.

Next we want to bound $\mathbb{E}[m_T]$ following the similar idea as in the original UCB paper.

$$\mathbb{E}\left[m_T(i)\right] = m + \sum_{t\in\mathcal{I}, t>\mathcal{I}[m]} Prob\left[(i_t = i) \quad and \quad m_{t-1} \geq m\right]$$
$$\leq m + \sum_{t\in\mathcal{I}, t>\mathcal{I}[m]} \underbrace{Prob\left[UCB_t(i) > UCB_t(i^*) \quad and \quad m_{t-1} \geq m\right]}_{\text{TERM1}}$$

Here m can be any non-negative integer. In the later analysis, choice of m helps us to get a tighter bound.

Now we analyze the TERM1.

$$\text{TERM1} \leq Prob\left[UCB_t(i^*) < \mu(i^*)\right] + Prob\left[UCB_t(i) > \mu(i^*) \quad and \quad m_{t-1} > m\right]$$
$$\leq Prob\left[\mu(i^*) - \hat{\mu}_t(i^*) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i^*)}}\right]$$
$$+ Prob\left[\hat{\mu}_t(i) - \mu(i) \geq \Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \quad and \quad m_{t-1} > m\right]$$

First, observe that $Prob\left[\mu(i^*) - \hat{\mu}_t(i^*) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i^*)}}\right]$ has nothing to do with m, we can directly apply Lemma 1 to get upper the bound. So now we want to know what is the $[k_s, k_e]$. First, because we made K uniform explore rounds at beginning, so $n_{t-1}(i)$ should at least be 1. Then, because at time t there will $\lfloor (t-1-K)v \rfloor$ blocks and in each block we pull each arm at least once due to pre-scheduling, so $k_s = \lfloor (t-1-K)v \rfloor + 1$. Finally, because each arm will have been pulled at least k_s times, so $k_e = t - 1 - (K-1)k_0$. So the upper bound is

$$\sum_{k=k_s}^{t-(K-1)k_s-1} \frac{1}{T^2} \leq \sum_{k=1}^{T-K\lfloor (T-1-K)v \rfloor} \frac{1}{T^2}$$
$$\leq \frac{(1-Kv)}{T} + \frac{Kv + K^2v + 1}{T^2} \leq \frac{(1-Kv)}{T} + \mathcal{O}(\frac{K}{T^2})$$

Then we are going to deal with $Prob\left[\hat{\mu}_t(i) - \mu(i) \geq \Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \quad and \quad m_{t-1} > m\right]$. Again we want to use the Lemma 1, but we need to choose m at first. The reason we want to choose m is that we consider the first m epoch the bound will be very loose, so we can directly bound the probability by 1.

Notice we can easily make connections between $n_{t-1}(i)$ and m,

$$n_{t-1}(i) \geq \lfloor v(t-1-K) \rfloor + m_{t-1} + 1$$
$$\geq (m_{t-1} - \frac{1}{v} + K) * \frac{1}{\frac{1}{v} - K} + m_{t-1} + 1$$
$$\geq m_{t-1}(1 + \frac{v}{1-Kv}) \geq m(1 + \frac{v}{1-Kv})$$

Where the last inequality comes from $t > \mathcal{I}[m]$ which means $t \leq K + m + 1$.

By choosing $m = \lfloor \frac{16 \ln T}{\Delta_i^2} * \frac{1-Kv}{1-(K-1)v} \rfloor$,

$$\Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} = 4\sqrt{\frac{\ln T}{m(1+\frac{v}{1-Kv})}} - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}$$

Again replace the above result in the probability bound and use Lemma 1 as before, we get

$$\begin{aligned} &Prob\left[\hat{\mu}_t(i) - \mu(i) \geq \Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \quad and \quad m_{t-1} > m\right] \\ &\leq Prob\left[\hat{\mu}_t(i) - \mu(i) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}\right] \leq \frac{(1-Kv)}{T} + \mathcal{O}(\frac{K}{T^2}) \end{aligned}$$

Therefore, we conclude bound for $i \neq i^*$ that

$$\begin{aligned} \mathbb{E}[m_T(i)] &\leq \frac{16 \ln T}{\Delta_i^2}\left(\frac{1-Kv}{1-(K-1)v}\right) + \sum_{t\in\mathcal{I}, t>\mathcal{I}[m]}\left(2\frac{(1-Kv)}{T} + 2\mathcal{O}(\frac{K}{T^2})\right) \\ &\leq \frac{16 \ln T}{\Delta_i^2}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2(1-Kv)^2 + 2\frac{(1-Kv)}{Tv} + 2\mathcal{O}(\frac{K}{T}) \\ &\leq \frac{16 \ln T}{\Delta_i^2}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2\mathcal{O}(1) + 2\mathcal{O}(\frac{K}{T}) \end{aligned}$$

Now we can get the total regret is:

$$\begin{aligned} Reg_T &= \sum_{i\neq i^*} \Delta_i \mathbb{E}\left[m_T(i)\right] \\ &\leq \sum_{i:\Delta_i>0}\left[\frac{16 \ln T}{\Delta_i}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2(1-Kv)^2\Delta_i + 2\mathcal{O}(1)\right] + \mathcal{O}(1) \\ &\leq \sum_{i:\Delta_i>0}\left[\frac{16 \ln T}{\Delta_i}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2(1-Kv)^2\Delta_i\right] + \mathcal{O}(K) \end{aligned}$$

This bound is not always tight, because when $\Delta \to \mathcal{O}(\frac{1}{T})$ and $v \ll \frac{1}{K}$, this bound will become linear. Therefore, for any $\Delta \in [0, 1]$ we can further write that as

$$\begin{aligned} Reg_T &= \sum_{\Delta_i\leq\Delta} \Delta_i \mathbb{E}\left[m_T(i)\right] + \sum_{\Delta_i>\Delta} \Delta_i \mathbb{E}\left[m_T(i)\right] \\ &\leq \Delta\lceil(1-Kv)T\rceil + \sum_{\Delta_i>\Delta}\left[\frac{16 \ln T}{\Delta_i}\left(\frac{1-Kv}{1-(K-1)v}\right) + 2(1-Kv)^2\Delta_i\right] + \mathcal{O}() \end{aligned}$$

By choosing $\Delta = \sqrt{\frac{K \ln(T)}{T}}$, we got the worst case guarantee,

$$\begin{aligned} Reg_T &\leq T\Delta + \sum_{\Delta_i>\Delta}\left[\frac{16 \ln T}{\Delta_i} + 2\Delta_i\right] + \mathcal{O}(K) \\ &\leq \mathcal{O}(\sqrt{TK \ln T} + K \ln T) \end{aligned}$$

6.4 Proof for algorithm 2

6.4.1 Notations

Denote the distribution over arms at time t as p_t where $p_t(\text{argmax}_{i\in[K]}\, UCB_t(i)) = (1-Kv) + v$ and $p_t(i) = v, \forall i \in [K] \setminus \text{argmax}_{i\in[K]}\, UCB_t(i)$. And the best distribution as p^* where $p^*(i^*) = (1-Kv) + v$ and $p^*(i) = v, \forall i \in [K] \setminus i^*$.

6.5 Main Proof

First we rewrite this regret in the form of variable Δ_i and m_T,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{i_t\sim p^*}[r_t(i_t)]-\mathbb{E}_{i_t\sim p_t}[r_t(i_t)]\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{i_t\sim p^*}[\mu(i_t)]-\mathbb{E}_{i_t\sim p_t}[\mu(i_t)]\right]\\
&=\mathbb{E}\sum_{t=1}^{T}\left[(1-(K-1)v)\mu(i^*)+v\sum_{i\neq i^*}\mu(i)-p_t(i^*)\mu(i^*)-\sum_{i\neq i^*}p_t(i)\mu(i)\right]\\
&=\mathbb{E}\sum_{t=1}^{T}\left[(1-p_t(i^*))\mu(i^*)-\sum_{i\neq i^*}p_t(i)\mu(i)+v\sum_{a\neq i^*}\mu(i)-\mu(i^*)\right]\\
&=\mathbb{E}\sum_{t=1}^{T}\left[\sum_{i\neq i^*}p_t(i)\Delta_i-v\sum_{i\neq i^*}\Delta_i\right]\\
&=\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[\sum_{t=1}^{T}p_t(i)\right]-vT\sum_{i\neq i^*}\Delta_i\\
&=\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{i_t=i\}\right]-vT\sum_{i\neq i^*}\Delta_i\\
&=\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[n_T(i)\right]-vT\sum_{i\neq i^*}\Delta_i
\end{aligned}
$$

Notice here the expectation is regarding to the both the randomness of environment and learner's choice of i_t, which is a bit different from previous proof. The penultimate equality is due to $\mathbb{E}_{\text{leaner at t}}[\mathbf{1}\{i_t=i\}]=p_t(i)$ and the linearity of expectation.

Next we want to bound $\mathbb{E}\left[n_T\right]$ following the similar idea as in the original UCB paper.

$$
\begin{aligned}
\mathbb{E}[n_T(i)] &= n+\sum_{t=n+1}^{T}Prob\left[(i_t=i)\quad and\quad n_{t-1}>n\right]\ (n \text{ here is simply for analysis})\\
&\leq n+\sum_{t=n+1}^{T}\left[(1-Kv)*\underbrace{Prob(UCB(i)>UCB(i^*)\quad and\quad n_{t-1}(i)>n)}_{\text{Term1}}+v\right]
\end{aligned}
$$

Here n can be any non-negative integer. In the later analysis, choice of n helps us to get a tighter bound.

Now we analyze the TERM1 using the almost same technique as proof for algorithm 2

$$
\begin{aligned}
\text{TERM1} &\leq Prob\left[UCB_t(i^*)<\mu(i^*)\right]+Prob\left[UCB_t(i)>\mu(i^*)\quad and\quad n_{t-1}>n\right]\\
&\leq Prob\left[\mu(i^*)-\hat{\mu}_t(i^*)\geq 2\sqrt{\frac{\ln T}{n_{t-1}(i^*)}}\right]\\
&\quad+Prob\left[\hat{\mu}_t(i)-\mu(i)\geq\Delta_i-2\sqrt{\frac{\ln T}{n_{t-1}(i)}}\quad and\quad n_{t-1}>n\right]
\end{aligned}
$$

First, observe that $Prob\left[\mu(i^*) - \hat{\mu}_t(i^*) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i^*)}}\right]$ has nothing to do with n, we can directly apply Lemma 1 to get upper the bound. Again we want to know what is the $[k_s, k_e]$. Because there is no prescheduling here, so it is simply just $[1, t-K]$. So the upper bound is

$$\sum_{k=1}^{t-K} \frac{1}{T^2} < \frac{1}{T}$$

Then we are going to deal with $Prob\left[\hat{\mu}_t(i) - \mu(i) \geq \Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \quad and \quad n_{t-1} > n\right]$. Again we want to use the Lemma 1, but we need to choose n at first. The reason we want to choose n is that we consider the first n epoch the bound will be very loose, so we can directly bound the probability by 1. However, consider the extreme case where $v \to \frac{1}{K}$, so $(1 - Kv) \to 0$, so the choice of arm is totally random and has nothing to do with UCB algorithms, so we can simply choose $n = 0$ and all the probabilities will be naturally bounded by 1.

Therefore, here I compute two cases, $n = [\frac{16 \ln T}{\Delta_i^2}]$ and $n = 0$.

When $n = 0$, simply bound the probability by 1.

When $n = [\frac{16 \ln T}{\Delta_i^2}]$, observed that

$$\Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} = 4\sqrt{\frac{\ln T}{n}} - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}$$

We can again apply Lemma 1 as before and get

$$\begin{aligned} &Prob\left[\hat{\mu}_t(i) - \mu(i) \geq \Delta_i - 2\sqrt{\frac{\ln T}{n_{t-1}(i)}} \quad and \quad n_{t-1} > n\right] \\ &\leq Prob\left[\hat{\mu}_t(i) - \mu(i) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(i)}}\right] \leq \frac{1}{T} \end{aligned}$$

Therefore, combine the two cases, we conclude bound for $i \neq i^*$

$$\mathbb{E}[n_T(i)] \leq \min\left\{\frac{16 \ln T}{\Delta_i^2} + (1 - Kv), (1 - Kv)T\right\} + vT$$

Now we can get the total regret is:

$$\begin{aligned} Reg_T &= \sum_{i \neq i^*} \Delta_i \mathbb{E}\left[n_T(i)\right] - vT \sum_{i \neq i^*} \Delta_i \\ &\leq \sum_{i:\Delta_i > 0} \left[\min\left\{\frac{16 \ln T}{\Delta_i} + (1 - Kv)\Delta_i, (1 - Kv)\Delta_i T\right\}\right] \end{aligned}$$

This bound is not always tight, because when $\Delta \to \mathcal{O}(\frac{1}{T})$ and $v \ll \frac{1}{K}$, this bound will become linear. Therefore, for any $\Delta \in [0, 1]$ we can further write that as

$$\begin{aligned} Reg_T &\leq \sum_{\Delta_i \leq \Delta} \Delta_i \mathbb{E}\left[n_T(i)\right] + \sum_{\Delta_i > \Delta} \Delta_i \mathbb{E}\left[n_T(i)\right] - vT \sum_{\Delta_i > \Delta} \Delta_i \\ &\leq \Delta T + \sum_{i:\Delta_i > \Delta} \left[\min\left\{\left(\frac{16 \ln T}{\Delta_i} + (1 - Kv)\Delta_i\right), (1 - Kv)\Delta_i T\right\}\right] \end{aligned}$$

By choosing $\Delta = \sqrt{\frac{K \ln(T)}{T}}$, we got the worst case guarantee,

$$\begin{aligned} Reg_T &\leq T\Delta + \sum_{\Delta_i > \Delta} \left[\frac{16 \ln T}{\Delta_i} + 2\Delta_i\right] + \mathcal{O}(K) \\ &\leq \mathcal{O}(\sqrt{TK \ln T} + K \ln T) \end{aligned}$$

6.6 Proof for minimum pulling rate for Algorithm 1

Here we prove that for strict-rate-control algorithm, the pulling rate is always at least $v - \frac{1}{t}$.

At time t, the arm i will be pulled at least $\lfloor tv \rfloor$ times according to the pre-scheduling. So the pulling rate will be $\frac{\lfloor tv \rfloor}{t} \geq \frac{tv-1}{t} = v - \frac{1}{t}$.