Active Learning for Fair and Stable Online Allocations

Riddhiman Bhattacharya, Thanh Nguyen, Will Wei Sun, Mohit Tawarmalani
Purdue University

Abstract

We explore an active learning approach for dynamic fair resource allocation problems. Unlike previous work that assumes full feedback from all agents on their allocations, we consider feedback from a select subset of agents at each epoch of the online resource allocation process. Despite this restriction, our proposed algorithms provide regret bounds that are sub-linear in number of time-periods for various measures that include fairness metrics commonly used in resource allocation problems and stability considerations in matching mechanisms. The key insight of our algorithms lies in adaptively identifying the most informative feedback using *dueling upper and lower confidence bounds*. With this strategy, we show that efficient decision-making does not require extensive feedback and produces efficient outcomes for a variety of problem classes.

Keywords: Bandit algorithms; dynamic fair allocation; regret analysis; stable matching

1 Introduction

Ensuring fair and stable allocation of scarce resources is a fundamental challenge in a wide range of applications. Traditional literature assumes that information regarding agents' preferences, whether available centrally to the designer or held privately by the agents, is known before the allocation process (the mechanism). However, this assumption hinders application in practical settings where agents typically evaluate resources only after receiving or consuming them. Furthermore, such preference information is often noisy and expensive for the central designer to gather from all agents, thus complicating the implementation of traditional mechanisms.

Examples of domains where these challenges manifest include applications where geographical and time constraints impede information collection, such as distributing resources to food banks and providing humanitarian aid to disaster areas and war zones [1, 6]. Even in online marketplaces devoid of physical constraints, such as dating services and job matching, evaluating information and collecting data presents a formidable challenge. Participants in these systems often assess compatibility only after the job commences or partnership begins, revealing the limitations of relying on pre-established preferences. Additionally, platforms themselves must exert significant effort to gather feedback through surveys and other mechanisms.

Recent literature bridges this gap partially by learning noisy preferences as allocation decisions are made. This approach makes allocation processes more adaptable and efficient when the information is incomplete or dynamically changing. However, the current research typically assumes that input from *all* participants is available at each time-epoch of the allocation process [12, 41, 27, 30, 14]. Since gathering information is costly and often practical considerations make it infeasible, assuming its availability overlooks

[1]Daniel School of Business, Purdue University, USA, bhatta76@purdue.edu. .

the possibility of designing efficient algorithms that operate with limited feedback and the accompanying analysis fails to illuminate which feedback is crucial for efficient design.

Our paper contributes on three fronts. First, we introduce a deliberate constraint on feedback, restricting it to a single agent or a limited number of agents per period instead of allowing input from all agents. Second, our paper makes a methodological contribution by developing a versatile framework that applies to both max-min/min-max envy scenarios and stable matching problems. The versatility of our approach underscores its adaptive and comprehensive nature, demonstrating that it is effective across diverse problem domains. Third, a key theoretical contribution of the paper is that, despite restricted feedback, our algorithms do not sacrifice regret significantly while addressing fairness and stability concerns. Our approach hinges on an active-learning procedure that carefully selects the agent from whom to gather feedback, ensuring its effectiveness in the allocation process. In the following, we describe a series of problems, with increasing degrees of complexity, and briefly describe our solutions. Figure 1 provides an outline of our paper. The simplest version we treat involves unit-demand agents seeking to consume

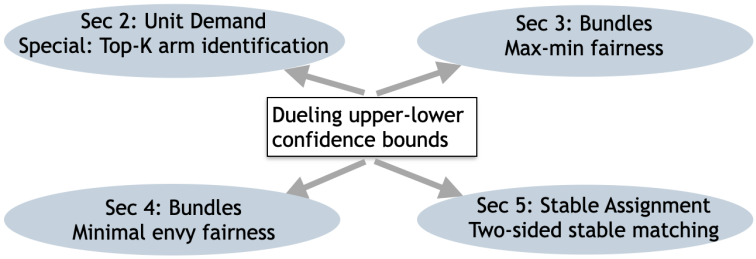


Figure 1: Outline of the four interconnected problems addressed in our paper.

single, indivisible items from a diverse set of resources. Minimax fairness focuses on the lowest reward any agent receives in the allocation, and our algorithm aims to maximize this reward. The online variant assumes that the reward matrix is unknown, which must be learnt during the allocation process. We additionally impose that feedback should be collected from one agent at each time-period. As specified, the problem adds to the growing literature on multi-armed bandit problems. A special case of interest is where rewards are agent-independent, and, in this setting, the problem reduces to the classical problem of finding the top K arms. Even for this special case, which is well-studied [13, 31, 39, 19, 42], limited feedback is new since earlier studies uniformly assume that feedback from all K arms is available at each time. We introduce Algorithm 1 based on the innovative concept of dueling upper-lower confidence bounds (dueling-ULCB). In this approach, we select the allocation based on the upper confidence bound (UCB) but choose the feedback based on the lower confidence bound (LCB). Out of various natural ideas for selecting feedback, this method proves to be the most effective, resulting in an algorithm with a sub-linear regret. Furthermore, the intuition behind it facilitates extensions to more complex problems.

We extend the original problem setting in three directions. First, Section 3 explores a bundle setting where each agent is allocated a set of items rather than just one item. This corresponds to extending the classical MAB setting to the combinatorial multi-armed-bandit (CMAB) setting [15, 16] where the agents at each epoch pull a set of arms (super arms) instead of one arm. The technical challenge this setting poses is that the decision and sample space may be exponentially sized in the number of arms. To avoid regret depending on this large sample space, our method utilizes LCB and UCB bounds for each individual good rather than treating super-arms individually. Consistent with limited feedback, throughout the paper, we are interested in the amount of feedback collected, and solve decision problems in each time-period using oracles which may have exponential complexity depending on the problem complexity.

The second extension relates to the objective. Instead of pursuing the max-min objective, Section 4 aims to find an assignment that minimizes the maximum envy between any pair of agents. The envy of agent i towards agent i' is defined as the gain in reward if agent i were to receive the bundle allocated to agent i'. The main challenge lies in the non-monotonic nature of the envy measure, requiring caution in the application of dueling ULCB. Instead, we use the the main insight from UCB and LCB to construct upper and lower estimates of the true envy and leverage these estimates in a dueling fashion. Our main results show that this approach allows us to identify allocations with optimal envy incurring sub-linear regret.

The third extension concerns stable matching, shifting the emphasis from maximizing an objective to the pursuit of a stable solution or determining its nonexistence. The key observation is that in many setting stability constraints are "local" constraints that involve a small number of agents. Therefore, in each period, the algorithm selects a constraint that violates stability the most to collect feedback. We furnish an algorithm that allows us to identify stable matchings in all but $O(\log T)$ epochs on average.

In each section, we assume computational oracles exist to solve certain subproblems. Our focus will remain on sampling complexity rather than computational efficiency. Even if the assumed oracles for computing optimal decision allocations in each period are precise, challenges still arise when determining which feedback to select. Our analysis can be readily extended to include approximation oracles. In the examples discussed in this work, we can regard these oracles (exact or approximate) as linear or integer programs that solve the static problem.

1.1 Related Works

Our paper contributes to the existing literature on online allocations by integrating two unique aspects. Firstly, we incorporate learning with noisy (bandit) feedback, which enhances the adaptability of our approach. Secondly, we impose strict constraints on the number of feedback instances. We elaborate on these differences in comparison to the three existing lines of work.

Online fair allocation with noiseless feedback: The online fair allocation, where items arrive dynamically and must be allocated to agents without revocation, has received extensive attention in the literature. Research has explored allocations that adhere to principles of fairness and efficiency [1, 2, 3, 4, 5, 10]. Recent efforts have considered online max-min fair allocations in adversarial settings [24, 17], and have developed approaches aimed at maximizing welfare or minimizing envy with full or partial information [33, 11, 7, 9, 8]. All the existing body of work assumes noiseless utility, where the true utility is observed in each allocation instance. In such cases, no learning mechanisms are involved, and the focus remains on efficiently achieving online fair allocation objectives. However, in many real examples, the precise observation of utility is not always possible, necessitating the handling of noisy feedback regarding the utility of the item received by the agent.

Online fair allocation with bandit learning: Recent developments in online fair allocation have increasingly emphasized the utilization of bandit learning [12, 41, 27]. These approaches are designed to tackle the challenge when the central planner does not have precise knowledge of agents' utilities. Diverging from traditional online algorithms, these approaches rely on noisy, estimated utilities obtained after item allocations. Moreover, they integrate the concept of UCB from multi-armed bandit problems to enhance the efficiency and fairness of online allocation processes. However, much of the current research assumes access to input from all participants at each allocation time-epoch, which may not be practical due to limitations on information gathering or the high cost associated with collecting feedback. To address it,

we propose active learning strategies aimed at gathering the most informative feedback from a single agent (or few agents) per step. This strategic adaptation is necessary due to the limited feedback available at each time instance, rendering existing algorithms based solely on UCB techniques inadequate.

Online stable matching with bandit learning: Recent research has applied bandit learning techniques to the domain of online stable matching, effectively framing the two-sided competing matching problem within a sequential decision-making framework [30, 14, 34, 22, 28, 29, 35]. For instance, [30] tackle the centralized multi-agent multi-armed competing bandit problem, where arms' preferences over agents are known, while agents' preferences over arms need to be learned from historical data. This work marks one of the pioneering efforts in online stable matching, considering the scenario where agents learn their preferences through bandit techniques. Subsequent studies have explored various aspects of bandit learning in online stable matching, such as handling unknown true preferences from both sides [14], episodic reinforcement learning settings [34], incorporating contextual information [29], and time-varying matching [35]. However, existing works typically assume observable noisy feedback from all matched pairs at each time instance, enabling the application of UCB-type or simple ETC-type algorithms. In our setting, feedback collection is costly and only one feedback is observed at each time. This motivates us to devise new dueling-type algorithm to incorporate both UCB and LCB of the estimated utility to address these challenges.

2 Max-min Fairness for Unit Demand Agents

In this section, we explore the proposed fair allocation framework tailored to unit demand agents, wherein a central platform distributes indivisible resources among agents who each consumes a single unit of a good. Following sections will expand this framework to encompass more diverse scenarios.

In the online setting, at each epoch $t = 1, 2, \cdots, T$, the platform allocates K out of the N goods (represented by $\mathcal{N} = \{1, 2, 3, \cdots, N\}$) among the K agents (represented by $\mathcal{K} = \{1, 2, \cdots, K\}$). Upon receiving a good i, the agent j receives a noisy reward that is given as $X_{ij}(t) = \mu_i^j + \epsilon_{ij}(t)$, where μ_i^j is the unknown true reward of agent $j \in \mathcal{K}$ being assigned an item $i \in \mathcal{N}$. The $\epsilon_{ij}(t)$ are subgaussian distributions with mean 0 and variance σ^2 for some $\sigma > 0$, and are independent and identically distributed across t. The goal is to assign at most one good to each agent in order to achieve max-min fairness, ensuring that the reward of the lowest-rewarded agent is maximized. Specifically, let $\mathcal{M}$ denote the set of all possible allocations of agents to goods, represented by $\phi : \mathcal{K} \to \mathcal{N}$. When the true reward μ_i^j is known, the optimal max-min objective is

$$\mu_{\text{maxmin}} = \max_{\phi \in \mathcal{M}} \min_{j \in \mathcal{K}} \mu^j_{\phi(j)}. \tag{2.1}$$

To evaluate the performance of a given allocation policy that allocates a good $\phi_t(j)$ to the agent j at time t, we employ the expected cumulative regret. This metric measures the accumulated difference between the objective from the proposed policy and the optimal max-min objective. Specifically, the expected cumulative regret over time horizon T is

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu_{\text{maxmin}} - \min_{j \in \mathcal{K}} \mu^j_{\phi_t(j)}\right)\right]. \tag{2.2}$$

The goal is to design an online allocation policy to minimize this expected cumulative regret. The difficulty lies in the combination of an unknown true reward that needs to be learned gradually over time and the constraint of a single feedback in each period. These challenges underscore the need for developing a new method for fair online allocation.

2.1 A Special Example: Top K-Arm Identification

To better understand the challenge and the intuition behind our approach, we start with a special case of this problem when the rewards are agent-independent, i.e., $\mu_i^j = \mu_i$ for all $j \in \mathcal{K}$. In this case, we can conceptualize the problem in a multi-arm bandit (MAB) setting. Each good i corresponds to an arm with an unknown reward μ_i, and the max-min allocation problem reduces to identifying the set of the best K arms. When $K = 1$, the problem reduces to the classical MAB problem, where a single arm is selected at each instance to identify the arm with the highest reward. One of the most widely employed algorithms in this context is the Upper Confidence Bound (UCB) algorithm, which operates on the principle of optimism in the face of uncertainty, pulling the arm that has the largest UCB [25]. Specifically, let's define $T_i(t)$ as the number of times an arm i has been pulled till time t, and additionally let $\hat{\mu}_i(t) = \sum_{l=1}^{T_i(t)} X_i(l)/T_i(t)$ denote the average of the collected noisy rewards obtained by pulling arm i. At time t, the UCB for arm i is defined as

$$\bar{\nu}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}. \tag{2.3}$$

Here, $\sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$ serves as a bonus term, ensuring that the algorithm explores the arms optimistically in the face of uncertainty and α is a constant larger than 2 [25]. We implicitly assume $\alpha > 2$ and $T > N$ in all our theoretical analysis and fix $\alpha = 3$ in all experiments. The UCB algorithm, at each time instance t, selects the i with the highest $\bar{\nu}_i(t)$ value. This algorithm has been analyzed extensively and it is known that each sub-optimal arm is pulled $O(\log T)$ times which gives an $O(N \log T)$ cumulative regret over total time horizon T [25].

For a general value of $K > 1$, previous work has explored the selection of the K best arms [39, 23, 43]. However, these studies necessitate pulling all K arms in each period. We constrain the feedback so that only one arm can be pulled. This introduces a new challenge, and to our knowledge, prior work does not specifically address this constraint. To demonstrate the difficulty, consider a simple MAB setting with 3 arms and the objective is to find the second best arm. If we use an analogue of the UCB algorithm here, i.e., find all the UCB estimates of the arms and choose the second highest UCB, then the algorithm fails to find the second best arm. This occurs because the error added to UCB increases if that arm is not explored. Consider the true arm ordering as $\mu_1 > \mu_2 > \mu_3$. It's possible that at a certain point t, the order of UCB does not align with the order of the real rewards, leading to a scenario such as $\bar{\nu}_2(t) > \bar{\nu}_1(t) > \bar{\nu}_3(t)$. If the algorithm continues exploring arm 1, $\bar{\nu}_3(t)$ and $\bar{\nu}_2(t)$ will keep increasing. This shall result in the ordering $\bar{\nu}_2(t) > \bar{\nu}_3(t) > \bar{\nu}_1(t)$. The algorithm then starts exploring arm 3, and keeps continuously exploring arms 1 and 3, while arm 2 remains unexplored. As a result, the algorithm fails to converge to the correct solution. Figure 3 in the simulation section shows that the cumulative regret of this "Second Best UCB" algorithm is linear in a MAB simulation with 3 arms.

To fix this, our proposal is to incorporate the Lower Confidence Bound (LCB) given as

$$\underline{\nu}_i(t) = \hat{\mu}_i(t) - \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}. \tag{2.4}$$

One notes that the LCB estimate has to be used in a correct fashion for the idea to work. An example of an incorrect usage of LCB is to select the arm with the K-th highest LCB estimate. This approach also

fails. Consider, again, the arm orderings as $\mu_1 > \mu_2 > \mu_3$. Then it is possible that at a certain point we shall have $\bar{\nu}_2(t) > \bar{\nu}_1(t) > \bar{\nu}_3(t)$ and $\underline{\nu}_2(t) > \underline{\nu}_3(t) > \underline{\nu}_1(t)$. In this case the algorithm explores arms 1 and 3, while never exploring arm 2. Therefore this approach also fails.

The key idea to remedy such issues is a new procedure called Dueling ULCB. It first identifies the arms with the top K UCB estimates and then selects the arm with lowest LCB estimate among the selected K arms. Our novel method effectively mitigates the challenges encountered in both of the aforementioned scenarios. Assume that we only select arms 1 and 3 for exploration. Then, for sufficiently large $\bar{t}$, we will have $\bar{\nu}_1(\bar{t}) > \bar{\nu}_3(\bar{t})$ and $\underline{\nu}_1(\bar{t}) > \underline{\nu}_3(\bar{t})$ as the UCB and LCB values under continuous exploration shrink to the true estimates. Moreover, $\underline{\nu}_1(\bar{t}) > \underline{\nu}_2(\bar{t})$ since $\underline{\nu}_2(\cdot)$ is a decreasing function while arm 2 is not pulled. Also, with sufficient pulls of arm 3, $\bar{\nu}_2(\bar{t}) > \bar{\nu}_3(\bar{t})$. This implies that arms 1 and 2 are identified by UCB, and among them arm 2 has a lower LCB which leads it to be being pulled. Among many other natural considerations, we show that this idea is highly effective, enabling us to address a wide range of problems. As shown in Figure 3, our approach demonstrates a substantial improvement over the "Second Best UCB" algorithm.

2.2 Our Algorithm and Regret Analysis

Now, we revisit the general fair online allocation problem, where we can represent each agent-good pair as an arm with μ_i^j for $j = 1, 2, \cdots, K$ and $i = 1, 2, \cdots, N$ as the true reward. The UCB and LCB estimates for the reward agent j receives from good i are given as

$$\bar{\nu}_i^j(t) = \hat{\mu}_i^j(t) + \sqrt{\frac{1}{T_i^j(t-1)} 2\sigma^2 \log t^\alpha}$$
$$\underline{\nu}_i^j(t) = \hat{\mu}_i^j(t) - \sqrt{\frac{1}{T_i^j(t-1)} 2\sigma^2 \log t^\alpha},$$

where $T_i^j(t)$ is the number of times that good-agent pair (i, j) has been chosen up to period t, and $\hat{\mu}_i^j(t) = \sum_{K=1}^{T_i^j(t)} X_{ij}(K)/T_i^j(t)$ is the estimate of the true reward.

Using the idea of dueling, we exploit the UCB and the LCB estimates to present an online algorithm which identifies the max-min allocation in most iterations. To achieve this goal, we need to balance between the need to acquire more knowledge about the reward distributions of each of the arms (exploration) and the need to estimate the max-min reward based on its current knowledge (exploitation). Before presenting our general Dueling ULCB algorithm in Algorithm 1, we introduce two oracles $\tilde{O}_1$ and $\tilde{O}_2$ that solve the corresponding static problems. Specifically, given any matrix $\mathbf{X} \in \mathbb{R}^{K \times N}$, $\tilde{\mathcal{O}}_1$ returns $\phi^* \in \arg\max_{\phi \in \mathcal{M}} \min_{j \in \mathcal{K}} x_{j\phi(j)}$, which is the assignment that gives the max-min allocation when $x_{ji} = \mu_i^j$. This problem can be solved via linear programming [20]. The second oracle $\tilde{\mathcal{O}}_2$ returns the minimum of a given a set of numbers. In our algorithm, we solve the max-min problem via $\tilde{\mathcal{O}}_1$ using UCB values of the arms and identify the minimum via $\tilde{\mathcal{O}}_2$ using the LCB values of the arms. In this work, we do not consider the algorithmic complexities of the oracles, but rather focus on the learning algorithm assuming the existence of such oracles. This is justified in our setting as we consider feedback to be costly and do not concern ourselves with the computational work required to solve the problems at each epoch. Although we do not detail here, approximation algorithms can be used to substitute our oracles with a corresponding loss of efficiency in regret. For each $j \in \mathcal{K}$, denote by $\bar{\boldsymbol{\nu}}^j(t) = (\bar{\nu}_1^j(t), \bar{\nu}_2^j(t), \bar{\nu}_3^j(t), \cdots, \bar{\nu}_N^j(t))$ as the vector of all UCB estimates of the goods for agent j. Algorithm 1 starts by pulling each arm once. Following this, the algorithm sequentially estimates the UCB and LCB values for each arm. At each epoch, the

algorithm uses the UCB estimates of the arms and the first oracle to compute the max-min allocation. Finally, the algorithm explores (seeks feedback from) the arm in the max-min allocation with the lowest LCB estimate.

Algorithm 1: Dueling ULCB Algorithm

Input K, α, σ^2.
for $t = 1, 2, \cdots, N \times K$
Pull each arm once.
Update:
$T_i^j(t)$, $\hat{\mu}_i^j(t)$, $\bar{\nu}_i^j(t)$ and $\underline{\nu}_i^j(t)$.
for $t = N+1, \cdots, T$ **do**:
UCB: $\bar{\nu}_i^j(t) = \hat{\mu}_i^j(t) + \sqrt{\frac{1}{T_i^j(t-1)} 2\sigma^2 \log t^\alpha}$
and
LCB: $\underline{\nu}_i^j(t) = \hat{\mu}_i^j(t) - \sqrt{\frac{1}{T_i^j(t-1)} 2\sigma^2 \log t^\alpha}$
Identify max-min allocation, ϕ_t using UCB values, i.e., $\phi_t = \tilde{\mathcal{O}}_1(\bar{\boldsymbol{\nu}}^1(t), \bar{\boldsymbol{\nu}}^2(t), \cdots, \bar{\boldsymbol{\nu}}^N(t))$.
Select the arm with the lowest LCB from the selected allocation as $I_t = \tilde{\mathcal{O}}_2(\phi_t, \underline{\boldsymbol{\nu}}(t))$.
Pull the arm I_t and **Output** (ϕ_t, I_t).

Next we study the theoretical properties of the proposed Dueling ULCB algorithm. We show that it indeed chooses the correct max-min allocation in most cases. The regret defined in (2.2) thus reduces to $R_T = \mathbb{E}\big[\sum_{t=1}^T \mu^{j^*}_{\phi^*(j^*)} - \min_{j\in\mathcal{K}} \mu^j_{\phi_t(j)}\big]$ where ϕ^* and j^* are a max-min allocation and an agent who receives the max-min allocation respectively and ϕ_t is the allocation chosen using Algorithm 1. Note that the regret is not defined with respect to the revealed arm, but with respect to the allocation ϕ_t. Observe that the regret is non-negative for each t since $\min_{j\in\mathcal{K}} \mu^j_{\phi_t(j)} \leq \mu^{j^*}_{\phi(j^*)}$, as at least one of the chosen agents receives a suboptimal allocation compared to j^* in the optimal solution. The main rationale for this choice of regret is that we wish to select the correct set by revealing only one arm. To establish our main results, we need to state an assumption on the true rewards. Define

$$\Phi^* = \left\{ \phi \in \mathcal{M} : \min_{j\in\mathcal{K}} \mu^j_{\phi(j)} = \mu_{\text{maxmin}} \right\}$$

as the set of allocations having the same true minimal allocation which is the max-min objective.

Assumption 1. *For any $\phi \in \mathcal{M}\backslash\Phi^*$, there exists a gap $\Delta_{\min} > 0$ such that*

$$\mu_{maxmin} - \min_{j\in\mathcal{K}} \mu^j_{\phi(j)} > \Delta_{\min}.$$

Furthermore, for any $i_1, i_2 \in \mathcal{N}$ and $j_1, j_2 \in \mathcal{K}$, one has

$$\left|\mu^{j_1}_{i_1} - \mu^{j_2}_{i_2}\right| \leq \Delta_{\max}.$$

Assumption 1 serves as a measure of identifiability for the max-min allocation. It's worth noting that this assumption is a generalization of the gap assumption in classical MAB setting [26], which assumes a gap between the top arm and all other arms. In our case, we posit a gap between our desired set of allocations (which may not be unique) and the remaining allocations. Moreover, Assumption 1 also requires that the differences between true rewards are upper bounded. This is a minor assumption and can be satisfied when the true reward is bounded.

THEOREM 1. *Under Assumption 1, the expected cumulative regret of Algorithm 1 satisfies*

$$R_T \leq 3\,\Delta_{\max}\,N\,K\left[\frac{\left(\sqrt{2\alpha}+2\right)^2\,\sigma^2}{\Delta_{\min}^2}\,\log T + 2\,\frac{\alpha-1}{\alpha-2}+2\right].$$

Next we discuss the regret bound with respect to a few key terms. First, the regret bound inflates with a decrease in $\Delta_{\min}$ as it becomes harder to identify the correct allocations. Further note that as $\Delta_{\max}$ increases, the regret increases as we penalize more when an incorrect allocation is chosen. With respect to the time horizon, the regret is sub-linear with rate $O(\log T)$. This implies that our algorithm is able to select the correct max-min allocation in most iterations. It's worth noting that Algorithm 1 incorporates the LCB step, enabling us not only to identify the true allocation but also the arm that achieves the true max-min objective. Finally, it's useful to note that when there is only one agent and the platform's task is to assign the agent the best item, the problem simplifies to the classical MAB setting where the objective is to obtain the best reward. By substituting $K = 1$ in the regret bound in Theorem 1, we get the same regret $O(N \log T)$ as that for the classic MAB setting.

In contrast to the proof in the classic MAB, our setting with a general $K > 1$ presents two key differences: the regret is based on a set rather than a single arm, and there are constraints on how we can explore the arms, specifically in terms of allocations. These factors make the problem considerably more challenging, and we leverage the properties of UCB and LCB to address them. In our approach, the regret bound is established using the ranking, where we demonstrate that the rankings of the UCB, LCB, and the true values should align after a sufficient number of arm pulls.

The regret bound in Theorem 1 relies on the gap assumption and is known as the instance-dependent bound. Next we provide a regret bound that is instance-independent. To ease the presentation, we consider the case where the rewards are identical for all agents, i.e., $\mu_i^j = \mu_i$ for all $j \in \mathcal{K}$. Note that in this setting, the max-min allocation is the set of the top K arms, and the K-th reward is the max-min objective. We refer to the true top K set as G^*. In this case, the regret defined in (2.2) reduces to $R_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\min_{i\in G^*}\mu_i - \min_{j\in G_t}\mu_j\right)\right]$, where G_t is the top K arms identified at time t using Algorithm 1. In this setting, the oracle $\tilde{\mathcal{O}}_1$ simply outputs the order-K statistic and we only need one oracle in this case. Define $\Delta_i(K) = |\mu_{(K)} - \mu_i|$ and

$$G(\delta) = \left\{i \in \mathcal{N} :\, \mu_{(K)} - \mu_i > \delta\right\}.$$

PROPOSITION 2.1. *In the setting $\mu_i^j = \mu_i$ for all $j \in \mathcal{K}$, the regret of Algorithm 1 over horizon T satisfies*

$$\begin{aligned} R_T \leq\; & 2\sqrt{((N-K+1)^2-1)\;K\,8\,\sigma^2\alpha\,T\,\log T} \\ & + \sum_{i\in G(0)}\left(\frac{2\,K\,\Delta_i(K)\,\alpha}{\alpha-2} + \frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2}\right). \end{aligned}$$

Since the second term in Proposition 2.1 is independent of the time horizon T, the overall rate of the regret bound in Proposition 2.1 with respect to the time horizon is of the order $O(\sqrt{T\,\log T})$. Note that the regret is defined in terms of the set selected and not the arm pulled. Therefore, in the case the correct minimal arm is not pulled, the regret is inflated by a factor of $\sqrt{N}$ to uniformly account for all the cases when the explored arm is not in the optimal set. Further note that in the case $K = 1$, such inflation of $\sqrt{N}$ does not appear and hence our rate shall match that in the MAB [25]. Thus the inflation of regret occurs due to definition of the regret and the uniform bounding of it over all cases. Note that when $K = N$, the

regret is zero. This is expected as the set $G(0) = \{i : \mu_i < \mu_{\text{maxmin}}\}$ becomes smaller when K increases and eventually becomes empty when $K = N$.

The primary challenge in this problem lies in addressing the additional errors resulting from regret being calculated with respect to the set rather than the explored arm in the case where there is a significant difference between the arms. To address this, we divide the problem into two cases- where the pulled arm is the true minimum of the considered set, and the other where it is not.

3 Allocation of Bundles

In this section, we expand upon the fair allocation framework introduced for unit demand agents in Section 2 to encompass scenarios involving bundles of goods.

3.1 Max-Min Fairness

We consider a scenario with a set of agents, labeled $\mathcal{K} = \{1, 2, \cdots, K\}$ and a set of goods, labeled $\mathcal{N} = \{1, 2, \cdots, N\}$. A *bundle* is defined as a subset of goods. Each agent is allocated a bundle in such a way that for any two agents, their bundles are disjoint. The list of bundles assigned to the agents collectively is referred to as an *allocation*.

Our main objective is to identify an allocation that optimizes fairness by maximizing the least reward attained by any agent. This goal extends the principle of max-min fairness to the case of bundles, rather than individual goods outlined in Section 2. Similar to our previous setting, we assume active feedback framework and require sampling the reward of only one agent.

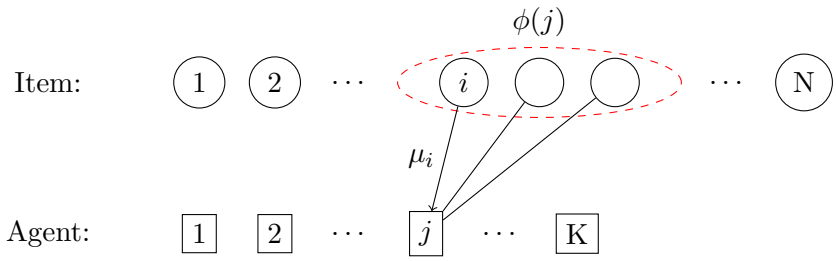


Figure 2: Illustration of an allocation with bundle $\phi(j)$ being allocated to agent j.

We assume, each agent $j \in \mathcal{K}$ has access to certain sets of bundles of goods which is denoted as $\mathcal{A}_j \subseteq 2^{\mathcal{N}}$. We call $\mathcal{A}_j$ as the set of feasible bundles for agent j. We assume $\emptyset \in \mathcal{A}_j$. The individual set $\mathcal{A}_j$ captures diverse geographical and technological constraints that vary among agents. We assume each good has a common true "quality" μ_i. Our result extends to the case where these qualities are agent-specific as well. For agent $j \in \mathcal{K}$, the reward for agent j, is given by a function

$$r^j : \mathbb{R}^N \times \mathcal{A}_j \to \mathbb{R}$$

In particular, given the true "quality" $\boldsymbol{\mu} = (\mu_1, .., \mu_N)$ and a feasible bundle S in $\mathcal{A}_j$, the true reward of agent j for receiving S is $r^j(\boldsymbol{\mu}; S)$. Note that in this context, we assume that the reward for a set S solely depends on the quality of items in S. However, to keep our notation simpler, we specify all qualities as arguments to the reward function $r^j(\boldsymbol{\mu}; S)$.

An allocation assigns a feasible bundle to each agent, ensuring that no item is assigned to more than one agent. In particular,

$$\mathcal{M} = \left\{ \phi : \mathcal{K} \to \bigcup_{j=1}^{K} \mathcal{A}_j : \phi(j) \in \mathcal{A}_j \text{ for all } j \in \mathcal{K}, \text{ and } \forall i \neq j, \phi(i) \cap \phi(j) = \emptyset \right\}$$

denote the set of all allocations. For every allocation $\phi \in \mathcal{M}$, $\phi(j)$ is the bundle assigned to j.[1]

In this setting we define the max-min problem as

$$\arg\max_{\phi \in \mathcal{M}} \min_{1 \leq j \leq K} r^j(\boldsymbol{\mu}; \phi(j)), \tag{3.1}$$

and denote $\text{opt}^{\text{maxmin}}$ as its oracle max-min objective when the true reward values are observed, i.e. $\text{opt}^{\text{maxmin}} = \max_{\phi \in \mathcal{M}} \min_{1 \leq j \leq K} r^j(\boldsymbol{\mu}; \phi(j))$. Note that there might be multiple solutions to the max-min objective. Thus, with some minor abuse of notation, as in Section 2, define $\Phi^* = \{\phi : \min_{j \in \mathcal{K}} r^j(\boldsymbol{\mu}; \phi(j)) = \text{opt}^{\text{maxmin}}\}$ as the optimal set of allocations. However, we posit that presenting any of these solutions is sufficient, and there is no imperative need to differentiate between them, as they all lead to the same reward outcome.

At every period t, if we choose to collect feedback of agent j for bundle S, we receive a noisy signal on the quality of good $i \in S$, $X_i(t) = \mu_i + \epsilon_i(t)$, and the corresponding noisy reward $r^j(\mathbf{X}(t); S)$. To evaluate the performance of any allocation policy, we compare its decision against an optimal benchmark that assumes full knowledge of the base reward vector and reward function. Thus, for any allocation policy that allocates $\phi_t(\cdot)$ at time t, its overall regret over time horizon T is quantified by

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(\text{opt}^{\text{maxmin}} - \min_{j \in \mathcal{K}} r^j(\boldsymbol{\mu}; \phi_t(j)) \right) \right]. \tag{3.2}$$

Considering that we're dealing with combinations of items, this scenario fits within the framework of the Combinatorial Multi-Armed Bandit (CMAB), which extends the conventional MAB model in Section 2. In CMAB, at each time step a player selects a combination of arms, termed as a super-arm, instead of a single arm. The reward from pulling a super-arm is determined by the rewards of its constituent individual arms. In the typical CMAB scenario, the objective is to identify the super-arm that yields the highest reward. In this case, algorithms leveraging UCB estimates of the base arms have been developed to achieve sub-linear regret [16]. However, our max-min fair allocation problem diverges from the standard CMAB framework: we are restricted to selecting super-arms that constitute a partition of $\mathcal{N}$, and our allocations consist of sets of super-arms rather than individual base arms. In our specific context, relying solely on algorithms utilizing UCB estimates for the base arms proves ineffective. This issue is explored in Section 2, where we present examples demonstrating the limitations of solely adjusting the UCB algorithm. To attain the desired results, it is imperative to also incorporate LCB estimates of the arms.

3.2 Our Algorithm and Regret Analysis

We present the algorithm to solve the online max-min allocation problem. In order to do this, we define the two oracles that are necessary for our algorithm. Define $\mathcal{O}_1$ as the oracle which takes a vector of N entries, the set of all super-arms, and the number of agents K and returns a max-min allocation. Namely, given any $\mathbf{x} \in \mathbb{R}^N$ and reward function r, $\mathcal{O}_1$ solves (3.1) where $\boldsymbol{\mu} = \mathbf{x}$. Additionally, we define another oracle, denoted as $\mathcal{O}_2$, which accepts a vector comprising N entries, which is a reward estimate for each item, along with an allocation, and outputs the agent with the lowest reward. Essentially, this oracle

operates similar to a quicksort algorithm. Again, as in Section 2, note that since we do not know the true rewards, i.e., $\boldsymbol{\mu}$, we must learn it through agent feedback. Denote the UCB vector of the base arms as $\bar{\boldsymbol{\nu}}(t) = (\bar{\nu}_1(t), \bar{\nu}_2(t), \cdots, \bar{\nu}_N(t))$ and the LCB vector of the base arms as $\underline{\boldsymbol{\nu}}(t) = (\underline{\nu}_1(t), \underline{\nu}_2(t), \cdots, \underline{\nu}_N(t))$. Our Dueling Max-Min ULCB is shown in Algorithm 2.

Algorithm 2: Dueling Max-Min ULCB Algorithm

Input: K, σ, α.
for: $t = 1, 2, \cdots, N$;
Pull each arm.
Update: $T_t(i), \hat{\mu}_i(t), \bar{\nu}_i(t), \underline{\nu}_i(t)$.
for $t = N+1, \cdots, T$ **do**:
UCB: $\bar{\nu}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$.
and
LCB: $\underline{\nu}_i(t) = \hat{\mu}_i(t) - \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$.
$(\phi_t, j_t) = \mathcal{O}_2(\underline{\boldsymbol{\nu}}(t), \mathcal{O}_1(\bar{\boldsymbol{\nu}}(t), \bigcup_{i=1}^{K} \mathcal{A}_j, K), K)$.
Output and Pull (ϕ_t, j_t) for $t = N+1, 2, \cdots, T$.

The algorithm operates as follows: initially, all base arms are pulled at least once. Subsequently, at each time step, the UCB and LCB estimates of the arms are updated. Based on these, the oracle $\mathcal{O}_1$ is used to obtain the max-min allocation and the oracle $\mathcal{O}_2$ is used to decide the arm to pull for collecting feedback. Our algorithm progressively transitions from exploration to exploitation. Initially, there may be incorrect allocations due to inadequate exploration. However, as time passes, the number of mistakes decreases, eventually resulting in correct identifying the max-min allocation.

To theoretically substantiate this, we next show that our algorithm achieves a sub-linear regret bound. Define $\phi^* = \mathcal{O}_1(\boldsymbol{\mu}, \bigcup_{i=1}^{K} \mathcal{A}_j, K)$ and $j^* = \mathcal{O}_2(\boldsymbol{\mu}, \phi^*, K)$ as the true max-min allocation and the agent receiving the max-min objective, respectively. Further note that, in this setting, the regret as defined in (3.2) reduces to

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(r^{j^*}(\boldsymbol{\mu}; \phi^*(j^*)) - \min_{1 \le j \le K} r^j(\boldsymbol{\mu}; \phi_t(j)) \right)\right]$$

where $\phi_t(\cdot)$ is the allocation chosen at time t. Since both j^* and $\phi(j^*)$ are optimal, this regret is always positive. It's important to note that this regret mirrors the definition of regret in the unit demand case. It's defined in relation to the true minimum of the selected allocation, rather than the revealed arm.

We state our assumptions for the main results of this section.

Assumption 2. *For any agent $j \in \mathcal{K}$ and any bundle $S \in \mathcal{A}_j$, there is some positive constant c such that $\left| r^j(\boldsymbol{\mu}; S) - r^j(\boldsymbol{\nu}; S) \right| \le c \sum_{i \in S} |\mu_i - \nu_i|$ for any two vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in $\mathbb{R}^N$. Furthermore, if $\mu_i \le \nu_i$ for all $i \in S$, one has $r^j(\boldsymbol{\mu}; S) \le r^j(\boldsymbol{\nu}; S)$, where μ_i and ν_i represent the i-th element of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively.*

Assumption 2 implies that the total reward increases when the reward for an item increases while other rewards do not decrease. Also a change in the reward for each agent is bounded with change in the rewards of base items. Such assumptions are common in the literature on combinatorial multi-armed bandits [16].

Next, we state one of our assumptions for the analysis in this setting which is akin to the gap assumption in the unit demand case.

Assumption 3. *For any $\phi_1 \in \Phi^*$ and $\phi_2 \in \mathcal{M}\backslash\Phi^*$, there exists $\tilde{\Delta}_{\min} > 0$ such that*

$$\left(\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_1(j)) - \min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j))\right) > \tilde{\Delta}_{\min}.$$

Further, for any ϕ_1, $\phi_2 \in \mathcal{M}$ and pair $i, j \in \mathcal{K}$, $i \neq j$,

$$\left|r^j(\boldsymbol{\mu};\phi_1(j)) - r^i(\boldsymbol{\mu};\phi_2(i))\right| \leq \tilde{\Delta}_{\max}.$$

One may note that for any $\phi_1 \in \Phi^*$, $\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_1(j)) = \text{opt}^{\text{maxmin}}$. Therefore Assumption 3 states that the optimal and the sub-optimal allocations are separated by some quantity $\tilde{\Delta}_{\min}$ in terms of the max-min objective. In essence, the optimal set of allocations are identifiable or discernible from the sub-optimal set of allocations by some identifiability constant in terms of the max-min objective. This identifiability assumption parallels Assumption 1 as stated in Section 2, in the bundle setting.

Theorem 2. *Under Assumptions 2-3, for Algorithm 2 , one has*

$$R_T \leq 3\,\tilde{\Delta}_{\max}\,N\left[\frac{\left(\sqrt{2\alpha}+2\right)^2 c^2\,N^2\,\sigma^2}{\tilde{\Delta}^2_{\min}}\,\log T + \frac{\alpha-1}{\alpha-2} + 2\right].$$

Theorem 2 establishes a sub-linear rate for the regret of Algorithm 2. Specifically, it makes at most $O(N^3\,\log T)$ errors within total time horizon T. Similar to the unit demand case, an increase in $\tilde{\Delta}_{\max}$ or a decrease in $\tilde{\Delta}_{\min}$ enlarges the regret bound. This is because, in the former scenario, errors are penalized more severely, while in the latter scenario, detection becomes more challenging due to a narrower gap.

The primary challenge in establishing this result lies in the reduction of this problem from super-arms to base arms. The combinatorial bandit can be conceptualized as a MAB where the super-arms are treated as base arms. However, this approach faces two primary challenges. Firstly, the number of super-arms can be exponential in the number of base arms, which would make the regret bound loose if we substituted N for the number of super-arms. Second, in the combinatorial setting, unlike in the unit demand setting in Section 2, the arms interact with one another via the reward function and feasibility considerations. To overcome these challenges in our proofs, we employ the novel idea that incremental knowledge of a few entries in low-dimensional quality vector, $\boldsymbol{\mu}$, improves our knowledge about rewards for all super-arms that contain those entries, and hence making the analysis tractable.

4 Minimal Envy Fairness

In this section, we illustrate the versatility of the proposed dueling UCB and LCB approach beyond the conventional max-min fair allocation problem. Here, we extend our methodology to address minimal envy [1, 2, 4] as an alternative fairness metric and tailor our method to identify online allocations that minimize envy.

The principle idea in envy is that each agent receives a satisfactory reward and does not desire another item. In this section, we define envy in the setting of our problem and propose an algorithm to minimize envy. As before, our set-up consists of K agents and N items. We start by defining the notion of envy among two agents. Given an allocation $\phi \in \mathcal{M}$ and any two agents $i, j \in \mathcal{K}$, the envy between the two agents i, j under allocation ϕ is given as

$$ev_{i\to j}(\boldsymbol{\mu},\phi) = \max\{r^i(\boldsymbol{\mu};\phi(j)) - r^i(\boldsymbol{\mu};\phi(i)), 0\}$$

where $ev_{i\to j}(\boldsymbol{\mu}, \phi)$ quantifies the extra reward agent i would attain upon acquiring the allocation originally designated for agent j. Further, for each allocation $\phi \in \mathcal{M}$, we define the envy of the allocation as

$$ev(\boldsymbol{\mu}, \phi) = \max_{i,j\in\mathcal{K}} ev_{i\to j}(\boldsymbol{\mu}, \phi) \tag{4.1}$$

which is the maximum envy between any two agents in the allocation ϕ. Our objective is to find the allocation that minimizes envy (4.1). In other words, we seek to find:

$$\phi^* = \arg\min_{\phi\in\mathcal{M}} ev(\boldsymbol{\mu}, \phi). \tag{4.2}$$

Let $\mathcal{E}^*$ represent the set of all allocations that achieve true minimal envy, or in other words, solutions of equation (4.2). To evaluate the policy that allocates ϕ_t at time t, we compare its envy objective to that of the optimal benchmark when the reward function and the true base rewards $\boldsymbol{\mu}$ are known. That is, for any $\phi^* \in \mathcal{E}^*$, we define the regret over time T as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(ev(\boldsymbol{\mu}; \phi_t) - ev(\boldsymbol{\mu}; \phi^*)\right)\right]. \tag{4.3}$$

By definition, this regret is always non-negative and is zero when $\phi_t \in \mathcal{E}^*$.

In practice, the true reward vector $\boldsymbol{\mu}$ is typically unknown, requiring us to learn it from agents' sequential feedback. Unlike the active feedback framework discussed earlier, where feedback was restricted to a single agent, in this minimal envy setting, we can collect feedback from two agents at each time point. This adjustment is essential for computing envy fairness, as it enables comparisons between two agents to ensure a fair assessment.

In line with our previous discussions, we could approach the problem within a bandit framework by employing UCB and LCB estimates of the base arms. However, the dueling ULCB approach used in previous sections is inadequate here. Blindly applying UCB and LCB values fails to accurately estimate the true envy and leads to erroneous identification. To address this issue, our core idea is to maintain an optimistic estimate of the true reward that steadily increases until exploration. Conversely, the LCB embodies the principle of pessimism, serving as an underestimation of the true reward that diminishes until exploration. Drawing on this principle, we recognize the need for both an optimistic and a pessimistic estimate of envy, each dynamically adjusting until exploration. Recall that we denote the UCB vector of the base arms as $\bar{\boldsymbol{\nu}}(t) = (\bar{\nu}_1(t), \bar{\nu}_2(t), \cdots, \bar{\nu}_N(t))$ and the LCB vector of the base arms as $\underline{\boldsymbol{\nu}}(t) = (\underline{\nu}_1(t), \underline{\nu}_2(t), \cdots, \underline{\nu}_N(t))$. For any allocation ϕ and agents $i, j \in \mathcal{K}$, we propose the following as the optimistic and pessimistic estimates of envy between agents i and j, respectively

$$\begin{aligned} ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi_t) &= \left(r^i(\bar{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\underline{\boldsymbol{\nu}}(t); \phi(i))\right)^+, \\ ev_{i\to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_t) &= \left(r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\bar{\boldsymbol{\nu}}(t); \phi(i))\right)^+, \end{aligned} \tag{4.4}$$

where $x^+ = \max\{0, x\}$. These values can be viewed as surrogates for the UCB and LCB estimates of envy. We designate the optimistic estimate as the upper estimate and the pessimistic estimate as the lower estimate of envy. Our key idea is to first use the lower estimate of the envy to solve

$$\arg\min_{\phi\in\mathcal{M}} \max_{i,j\in\mathcal{K}} ev_{i\to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_t),$$

which is a proxy of (4.2), and then use the upper estimate to find the pair of agents in an allocation with maximum envy, namely,

$$\arg\max_{i,j\in\mathcal{K}} ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi_t).$$

This approach reverses our previous application of upper and lower estimates (UCB and LCB) for addressing the max-min objective. The core principle is that objectives needing minimization should be approached with underestimation of true values, whereas maximization objectives warrant overestimation. We propose that this dueling dynamic of underestimation and overestimation will guide us in iteratively uncovering the true objective. Before presenting our algorithm, we assume the existence of two computational oracles $\mathcal{O}_1^E$ and $\mathcal{O}_2^E$ which solve (4.2) and (4.1) respectively given any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$, i.e.,

$$\tilde{\phi} = \arg\min_{\phi \in \mathcal{M}} \max_{i,j \in \mathcal{K}} ev_{i \to j}(x_1, x_2, \phi,) = \mathcal{O}_1^E(x_1, x_2, \bigcup_{j=1}^{K} \mathcal{A}_j, K) \tag{4.5}$$

$$(\tilde{i}, \tilde{j}) = \arg\max_{i,j \in \mathcal{K}} ev_{i \to j}(x_2, x_1, \phi_t) = \mathcal{O}_2^E(x_2, x_1, \tilde{\phi}, K). \tag{4.6}$$

Algorithm 3: Envy-combinatorial ULCB Algorithm

Input: K, σ, α.
for: $t = 1, 2, \cdots, N$;
Pull each arm.
Update: $T_i(t), \hat{\mu}_i(t), \bar{\nu}_i(t), \underline{\nu}_i(t)$.
for $t = N+1, \cdots, T$ **calculate**:
UCB: $\bar{\nu}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$.
and
LCB: $\underline{\nu}_i(t) = \hat{\mu}_i(t) - \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$.
$\phi_t = \mathcal{O}_1^E(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \bigcup_{i=1}^{K} \mathcal{A}_j, K)$.
$(i_t, j_t) = \mathcal{O}_2^E(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi_t, K)$.
Pull $(\phi_t(i_t), \phi_t(j_t))$ for $t = N+1, 2, \cdots, T$.
Output $(\phi_t, \phi_t(i_t), \phi_t(j_t))$ for $t = N+1, 2, \cdots, T$.

Our Envy-combinatorial ULCB algorithm is presented in Algorithm 3. After pulling each arm once, at each time instance we update the UCB and LCB values. Based on these, we solve (4.5) and (4.6) using $\mathcal{O}_1^E$ and $\mathcal{O}_2^E$ respectively with $\mathbf{x}_1 = \bar{\boldsymbol{\nu}}(t)$ and $\mathbf{x}_2 = \underline{\boldsymbol{\nu}}(t)$, then explore the outputs by pulling the two super-arms returned by the algorithm.

We next establish the upper bound of the cumulative regret for our algorithm. It's important to note that regret only arises when $\phi_t \notin \mathcal{E}^*$. To begin, we introduce an identifiability assumption.

Assumption 4. *For any $\phi_1 \notin \mathcal{E}^*$ and $\phi_2 \in \mathcal{E}^*$, there exists $\Delta_{e,\min} > 0$, such that*

$$ev(\boldsymbol{\mu}; \phi_1) - ev(\boldsymbol{\mu}; \phi_2) \geq \Delta_{e,\min}.$$

Further, for any ϕ_1, ϕ_2 and any pairs $(i,j) \neq (i',j') \in \mathcal{K} \times \mathcal{K}$, there exists $\Delta_{e,\max} > 0$ such that

$$\left| ev_{i \to j}(\boldsymbol{\mu}, \phi_1) - ev_{i' \to j'}(\boldsymbol{\mu}; \phi_2) \right| \leq \Delta_{e,\max}.$$

Assumption 4 is an identifiability assumption for the set of optimal envy allocations; akin to the previous assumptions in Sections 2 and 3, allowing us to discern an allocation as being optimal envy or not.

THEOREM 3. *Under Assumptions 2 and 4, we have*

$$R_T = O\left(\frac{N^3 \log T}{\Delta_{e,\min}^2}\right).$$

Theorem 3 implies that Algorithm 3 makes at most $O(N^3 \log T)$ errors in selecting the allocation with minimal envy. The challenges in the proof lie in the formulation of the problem, obtaining tail bounds and finally establishing the envy regret. Due to the need to account for comparisons between pairs of agents, envy is not necessarily a monotone function of the reward. This necessitates a more intricate analysis compared to previous formulations where the objective was more amenable to simpler methods. To address these challenges, we prove the validity of the constructed upper and lower estimates of the reward by deriving tail bounds for them. Moreover, we prove that the relative order between these estimates remains consistent after a substantial number of pulls.

5 Stable Assignment

In this section we apply our dueling ULCB framework to address more general problems of stability. The primary distinction in the stable matching problem, compared to the fair allocation discussed in the previous section, lies in the absence of goods. Instead, agents are paired together, and it becomes essential to account for the potential deviation of a subset of agents from the centralized solution.

We first present a framework that closely aligns with the one discussed in the preceding section, yet it is sufficiently versatile to encompass various stable matching models, such as college admissions [18, 38], ride-sharing [32, 40], and matching with couples [37, 36].

5.1 Model

We continue to use the notation outlined in previous sections, where $\mathcal{N}$ denotes the set of goods and $\mathcal{K}$ denotes the set of agents. Each agent j has a set of accessible bundles from $\mathcal{K}$, denoted by $\mathcal{A}_j$, along with a reward function given as $r^j : \mathbb{R}^N \times \mathcal{A}_j \to \mathbb{R}$ for $j = 1, 2, \cdots, K$. We further define the desired set of *feasible* assignments as

$$\mathcal{M}^* \subseteq \{\phi : \mathcal{K} \to \cup \mathcal{A}_j \mid \phi(j) \in \mathcal{A}_j\}.$$

$\mathcal{M}^*$ represents a *subset* of assignments with unspecified feasibility constraints, tailored to the specific problem. Allowing $\mathcal{M}^*$ to be arbitrary allows us to flexibly model a wide range of settings.

For example, in a marriage model (dating market), the set of agents consists of a union of two sets: $\mathcal{K} = \mathcal{G}_1 \cup \mathcal{G}_2$ where $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) is the set of men (resp. women) who are to be matched. The set of "goods" corresponds to all possible pairs of men and women to be $\mathcal{N} = \{(m, w), (w, m) : m \in \mathcal{G}_1 \text{ and } w \in \mathcal{G}_2\}$.[2] Each agent is interested to consume a single good. For an agent $j \in \mathcal{K}$, the set of accessible goods is $\mathcal{A}_j := \{(j, j') \mid j' \text{ is in the different gender group}\}$. A matching of men and women is a map $\phi : \mathcal{K} \to \mathcal{N}$ such that if $\phi(m) = (m, w)$ for $m \in \mathcal{G}_1$ then $\phi(w) = (w, m)$ for that particular $w \in \mathcal{G}_2$. Therefore, with this additional constraint, we define the set of all matching as:

$$\mathcal{M}^*_{\text{matching}} = \{\phi : \mathcal{K} \to \mathcal{N} \mid \text{if } \phi(a) = (a, b) \text{ then } \phi(b) = (b, a)\}.$$

Applying similar logic, the model can accommodate other matching problem variants. More specifically, the fundamental matching elements need not be confined to simple man-women matches; they can extend

to scenarios such as matching couples (where two students are paired with two hospitals) or many-to-one matching (where a group of students is matched to a single school).

Next we consider stability constraints. For example, in the marriage model, a matching is stable if there is no (m, w) who are not matched, but prefer to be together. To formulate this constraint in our general framework, we consider for each matching $\phi \in \mathcal{M}^*$, and a group of agents $L \subseteq \mathcal{K}$ there is a set of "deviating" assignment depending on L, ϕ, denoted as $\mathcal{M}^*|_{\phi,L}$, where

$$\mathcal{M}^*|_{\phi,L} \subseteq \{\phi' : L \to \cup_{j\in L}\mathcal{A}_j \mid \phi'(j) \in \mathcal{A}_j\}.$$

For instance, in a marriage model, given a matching function ϕ and two agents of different genders $L = \{m, w\}$ such that $\phi(m) \neq w$, the set of "deviating" assignments consists of a single element: $\phi'(m) = (m, w); \phi'(w) = (w, m)$. In alternative matching models, like matching with couples, a more comprehensive set of deviating agents needs to be considered. This set could include, for example, 2 students and 2 hospitals.

We now define the quality of the possible deviation which measures how much more reward an agent gets under the current matching compared with the deviation. Thus, the true benefit for an agent to stay in the current assignment given the deviation is computed as

$$g^j(\boldsymbol{\mu}; \phi \to \phi') := r^j(\boldsymbol{\mu}, \phi(j)) - r^j(\boldsymbol{\mu}, \phi'(j))$$

when the base rewards are $\boldsymbol{\mu}$ with the original matching being ϕ and the deviation being ϕ'.

If agent j experiences a negative benefit, it implies he is motivated to deviate from the current assignment ϕ. To measure the collective incentive of the entire agent group L to stick with the current matching rather than deviating to ϕ', we calculate:

$$g^L(\boldsymbol{\mu}; \phi \to \phi') := \max_{j\in L}\big\{g^j(\boldsymbol{\mu}; \phi \to \phi')\big\}.$$

A negative value for $g^L(\boldsymbol{\mu}; \phi \to \phi')$ indicates that the group L has a collective incentive to deviate and a positive value indicates that there exits at least one agent who has no incentive to deviate.

In the context of models such as the marriage model or matching with couples, we impose a restriction allowing only a limited number of agents to form a deviate coalition. Consequently, we assume that the size of the set L, denoted by $|L|$, is constrained such that $|L| \leq \kappa$, where κ is a small constant. For example in pairwise stable setups, κ is typically set to 2, while in the case of matching for couples, κ is set to 4.

Consider the following set that plays a key role in our problem:

$$\mathbb{F}(\boldsymbol{\mu}, \theta) = \left\{\phi \in \mathcal{M}^* \;\middle|\; g^L(\boldsymbol{\mu}, \phi \to \phi') \geq \theta \text{ for all } L \subseteq \mathcal{K}, |L| \leq \kappa, \phi' \in \mathcal{M}^*|_{L,\phi}\right\}.$$

The set $\mathbb{F}(\boldsymbol{\mu}, \theta)$ represents all matchings wherein any group of agents with at most κ members has an incentive of at least θ to resist changing their current assignment. We refer to this matching set as *θ-stable* and any element ϕ within it is termed a θ-stable matching. Additionally, when $\theta = 0$, we colloquially refer to the set as stable and its elements as stable matchings.

5.2 Learning Task and Regret Measures

Our objective is to determine whether the set of of stable matching $\mathbb{F}(\boldsymbol{\mu}, 0)$ is empty or not, and to select a stable matching if the set is non-empty without knowing the true $\boldsymbol{\mu}$. For this purpose, we pose this

question as a statistical hypothesis testing problem. Given *any* $\eta > 0$, we formulate the null and the alternate hypothesis as follows

$$H_0 : \mathbb{F}(\boldsymbol{\mu}, 0) = \emptyset; \quad H_a : \mathbb{F}(\boldsymbol{\mu}, \eta) \neq \emptyset.$$

The null hypothesis posits the absence of any stable matching, whereas the alternative hypothesis suggests the presence of a strong stable matching, wherein every deviation involves at least one agent with a strict preference to remain.

To study this hypothesis we consider this problem in an online setting where we aim to learn the stability set by gathering feedback from agents. Like in previous section, we are not allowed to collect feedback regarding the entire matching, but are restricted to at most κ agents in each period. In each period, there is an identical group of agents with unknown characteristics $\boldsymbol{\mu}$. Our algorithm tests the hypothesis using noisy estimates of $\boldsymbol{\mu}$ through an online decision rule δ_t, which accepts or rejects the null hypothesis at time instance t. Additionally, we aim to output a matching believed to be stable when the alternate hypothesis is true.

Since we are analyzing a hypothesis testing problem, there are two errors- the Type I and the Type II errors that need to be taken into consideration. Thus, we naturally have three different regrets that measure the goodness of our decision rule and the quality of our solution. If the decision rule is such that we reject the null hypothesis if $\delta_t = 1$ and do not reject it when $\delta_t = 0$ then the three regret measures are as follows:

$$\text{Type I error: } R_T^{H_0} = \mathbb{E}_{H_0}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t = 1\right\}\right], \quad \text{Type II error: } R_T^{H_a} = \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t = 0\right\}\right], \tag{5.1}$$

and the regret associated with not identifying a feasible solution when it exists:

$$R_T = \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right]. \tag{5.2}$$

The first two regrets measure the rate of the Type I and Type II errors and the third one measures the rate at which we identify a solution in the stable set. We shall develop an algorithm that identifies the correct decision and solution with sub-linear regret.

5.3 Algorithm

Note that to analyze this setting, we use the idea of dueling UCB and LCB in previous sections. For any ϕ, $L \subseteq \mathcal{K}$, with $\phi' \in \mathcal{M}^*|_{\phi,L}$, we define the upper estimate and the lower estimate of $g^L(\boldsymbol{\mu}, \phi \to \phi')$ over $[\bar{\boldsymbol{\nu}}, \underline{\boldsymbol{\nu}}]$ as

$$\text{upper} : g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') = \max_{j \in L} r^j(\bar{\boldsymbol{\nu}}(t), \phi(j)) - r^j(\underline{\boldsymbol{\nu}}(t), \phi'(j))$$

and

$$\text{lower} : g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') = \max_{j \in L} r^j(\underline{\boldsymbol{\nu}}(t), \phi(j)) - r^j(\bar{\boldsymbol{\nu}}(t), \phi'(j)),$$

where $\bar{\boldsymbol{\nu}}(t)$ and $\underline{\boldsymbol{\nu}}(t)$ are the UCB and LCB estimates of the true base rewards respectively at time t.

The decision function for the hypothesis is given as,

$$\begin{aligned}\delta_t^\epsilon &= 1 \text{ if there exists } \phi \text{ such that } \; g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \epsilon, \;\; \forall\, L \subseteq \mathcal{K} \text{ and } \phi' \in \mathcal{M}^*|_{\phi,L} \\ &= 0 \text{ otherwise.}\end{aligned}$$

Note that this decision rule is the same as the decision rule δ_t while defining the regret. We write it as δ_t^ϵ to emphasize its dependence on ϵ, which is a relaxation parameter for testing the hypothesis.

For a given ϕ, if $g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi')$ is at least ϵ for all $L \subseteq \mathcal{K}$ and $\phi' \in \mathcal{M}^*|_{\phi,L}$, we say that ϕ is estimated to be ϵ-stable over *all* reasonable values of $\boldsymbol{\mu}$. On the other hand, if the upper-estimates $g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi')$ are at least ϵ for $L \subseteq \mathcal{K}$ and $\phi' \in \mathcal{M}^*|_{\phi,L}$, we say that ϕ could be ϵ-stable over *some* reasonable value of $\boldsymbol{\mu}$. Now, we consider an oracle $\mathcal{O}_F$ which, given $\bar{\boldsymbol{\nu}}, \underline{\boldsymbol{\nu}}$ each in $\mathbb{R}^N$ and an $\eta \in \mathbb{R}$, solves the static stability problem, and if $\delta_t^\epsilon = 1$ returns a corresponding $\phi \in \mathcal{M}^*$. Recall that, by definition, this ϕ is estimated to be ϵ-stable over *all* reasonable values of $\boldsymbol{\mu}$. On the other hand, if $\delta_t^\epsilon = 0$ no such ϕ exists. Instead, the oracle checks if there is a $\phi \in \mathcal{M}^*$ that could be η-stable with $g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta$ for all $L \subseteq \mathcal{K}$, $\phi' \in \mathcal{M}^*|_{L,\phi}$ which we define as a solution being η-stable for some value of $\boldsymbol{\mu}$. If so, the oracle returns this ϕ. Otherwise, it returns an arbitrary $\phi \in M^*$. Regardless of whether our estimates find that ϕ could be η-stable or not, since $\delta_t^\epsilon = 0$, ϕ is not estimated to ϵ-stable over all reasonable values of $\boldsymbol{\mu}$ and, therefore, there exists an $L \subseteq \mathcal{K}$ and a $\phi' \in M^*|_{L,\phi}$ such that the lower estimate $g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') < \epsilon$. The oracle returns this set which provides evidence of why ϕ was not estimated to be ϵ-stable. We present our algorithm in Algorithm 4. Note that the selection of ϵ is fundamental to the algorithm and hence practitioners are advised to use multiple values to see which works best in practise.

Algorithm 4: Feasibility ULCB

Input: $\mathcal{N}, \mathcal{K}, \sigma, \alpha, \eta$.

for: $t = 1, 2, \cdots, N$;

Pull each arm.

Update: $T_i(t), \hat{\mu}_i(t), \bar{\nu}_i(t), \underline{\nu}_i(t)$.

for $t = N+1, \cdots, T$ **do**:

UCB: $\bar{\nu}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$

and

LCB: $\underline{\nu}_i(t) = \hat{\mu}_i(t) - \sqrt{\frac{1}{T_i(t-1)} 2\sigma^2 \log t^\alpha}$

$(\delta_t^\epsilon, \phi_t, L_t, \phi_t') = \mathcal{O}_F(\eta, \underline{\nu}_1(t), \underline{\nu}_2(t), \cdots, \underline{\nu}_N(t), \bar{\nu}_1(t), \bar{\nu}_2(t), \cdots, \bar{\nu}_N(t))$

Output ϕ_t **and Pull** $\phi_t(j), \phi_t'(j)$ for $t = N+1, 2, \cdots, T$, $j \in L_t$.

Next we study the theoretical properties of our proposed feasibility ULCB algorithm by deriving the upper bounds of the type-I and type-II errors in (5.1) and the regret in (5.2). We first list the main assumption for this section. Define, for each ϕ, the set $\mathcal{B}_\phi = \left\{(L, \phi') : g^L(\boldsymbol{\mu}, \phi \to \phi') < \eta\right\}$. Note that $\mathcal{B}_\phi$ may be empty for a particular ϕ.

Assumption 5. *There exists an $\eta > 0$ and $\Delta_{\mathcal{M}^*,q} > 0$ such that*

$$\inf_{\phi \in \mathcal{M}^* \setminus \mathbb{F}(\boldsymbol{\mu},\eta),\, (L,\phi') \in \mathcal{B}_\phi} \left(\eta - g^L(\boldsymbol{\mu}, \phi \to \phi'))\right) \geq \Delta_{\mathcal{M}^*,q} > 0.$$

Further, we assume $\eta - \Delta_{\mathcal{M}^*,q} < \epsilon < \eta$.

Assumption 5 implies that there exists a gap between the stable set and its complement. The reason for this is that in the stable set $g^L(\boldsymbol{\mu}, \phi \to \phi') \geq \eta$ for all L, ϕ' and on the complement of a stable set, we have $g^L(\boldsymbol{\mu}, \phi \to \phi') < \eta - \Delta_{\mathcal{M},q}$. This is akin to canonical hypothesis testing scenarios where there needs to be difference between the signals for a true discovery to be identified. As in previous sections,

this may again be viewed as an identifiability condition for the stable set. Note that if H_0 holds then $\sup_{\phi\in\mathcal{M}^*}\inf_{(L,\phi')\in\mathcal{B}_\phi} g^L(\boldsymbol{\mu},\phi\to\phi'))<0$.

We are now ready to state our main results, of which the proof is provided in the Appendix.

PROPOSITION 5.1. *Assume that Assumption 2 holds. The expected number of type-I errors made by Algorithm 4 satisfies*

$$R_T^{H_0}\le\frac{2\,N\,(\alpha-1)}{\alpha-2}.$$

Proposition 5.1 implies that our algorithm is expected to make no more than a constant number of errors in identifying the null hypothesis regardless of the time horizon as long as $T>N$.

THEOREM 4. *Let Assumptions 2 and 5 hold. Then, the type-II error of our algorithm satisfies*

$$R_T^{H_a}\le C_1^*(\kappa,\sigma^2,r,\alpha)\left(\Delta_{\mathcal{M},q}^{-2}+2\,(\eta-\epsilon)^{-2}\right)N^3\,\log T+N\,C_2^*(\alpha)$$

where $C_1^*(\kappa,\sigma^2,r,\alpha)$, $C_2^*(\alpha)$ are quantities independent of N,T.

Theorem 4 implies that our algorithm expects to make $O(N^3\log T)$ type-II errors within a total time horizon of T, thereby accurately identifying feasibility in most iterations.

THEOREM 5. *Under Assumptions 2 and 5, we have*

$$R_T\le\tilde{C}_1^*(\alpha,r)\left(\Delta_{\mathcal{M},q}^{-2}+(\eta-\epsilon)^{-2}\right)N^3\,\log T+N\,\tilde{C}_2^*(\alpha)$$

where $\tilde{C}_1^*(\alpha,r)$ and $\tilde{C}_2^*(\alpha)$ are independent of N and T.

Theorem 5 implies that our algorithm identifies a stable solution in most iterations, making at most $O(N^3\,\log T)$ errors within a total time horizon of T.

Taken together, these results show that, even though the feedback we collect is restricted to sets of agents of size κ, each of which participates by assessing their local stability, our algorithm correctly detects the existence/non-existence of a globally-stable solution and returns stable assignments in most time-periods, provided η-stable solutions exist. Note that the smaller the value of $\Delta_{\mathcal{M},q}$ is the harder the problem is as not only does the signal become hard to detect but also the range of values for ϵ decreases.

6 Simulations

In this section we present simulation experiments which bolster our theoretical understanding of the problem. We first present the example considered in Section 2.1 with $N=3$ goods and $K=2$ agents. The true arm rewards are $\mu_1=1,\mu_2=2,\mu_3=3$ respectively for the three arms and the observed rewards at time t are $X_i(t)\sim N(\mu_i,1)$ where $i=1,2,3$. Our objective is to select the second-largest arm, i.e., arm 2. Figure 3 illustrates a comparison of cumulative regrets between our proposed Dueling ULCB and two benchmark methods. The first benchmark selects the arm with the second-best UCB and is referred to as "Second Best UCB". The second benchmark method sequentially selects the arm with the best UCB and subsequently the second best UCB. This refers to the sequential version of existing top-k-arm selection algorithm [21]. We refer to this as "Sequential UCB". Evidently, the benchmark methods fail to converge to the correct solution, while our approach demonstrates a substantial improvement in regret. This indicates that in our considered limited feedback setting, only using UCB is insufficient and it is critical to incorporate both UCB and LCB for the active learning.

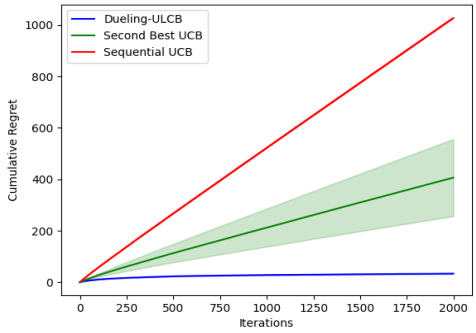
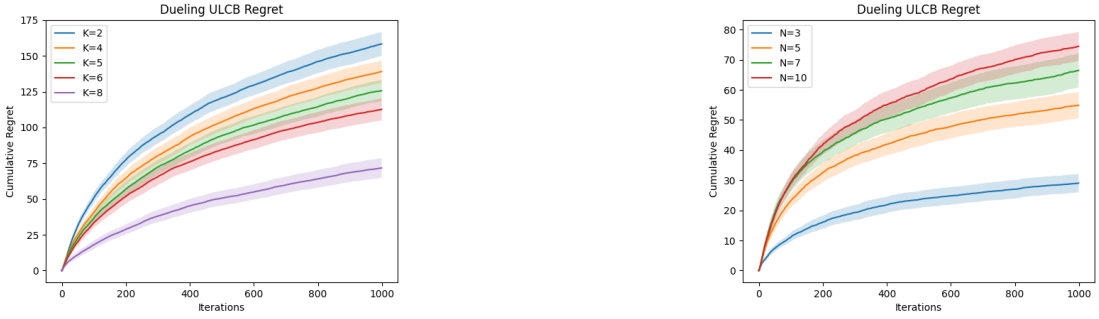


Figure 3: Comparison between the proposed Dueling ULCB and the two benchmark methods.

In fact, for any algorithm with the property that an incorrect arm is chosen at each instance with some fixed probability, i.e., $\mathbb{P}$(choosing some incorrect arm at time t) $\geq c$ where $c > 0$ is some constant, it would incur a linear regret. This is demonstrated in the Sequential UCB algorithm with 3 arms with $K = 2$ in Figure 3. Therefore, for an algorithm to exhibit sub-linear regret, the probability of selection of incorrect arms should decrease to 0 when time horizon increases.

Next we study the effects of K and N on the cumulative regret of the Dueling ULCB algorithm. We consider a MAB setting with the number of goods $N = 10$ and the number of agents $K = 2, 3, 5, 7, 8$. Further, we generate the reward for the i-th agent as $X_i(t) \sim N(\mu_i, 1)$, where $\mu_i = i$, $i = 1, 2, \cdots, N$. The arm differences are equal to 1 in this case. In Figure 4 we see that the cumulative regret is sub-linear with respect to the number of iterations. Further, in Figure 4a the cumulative regret of the algorithm shows that as K increases the regret decreases. Moreover, Figure 4b shows that the regret is increasing in N, which is expected as more exploration is needed for larger N. This result is also consistent with our theoretical findings.

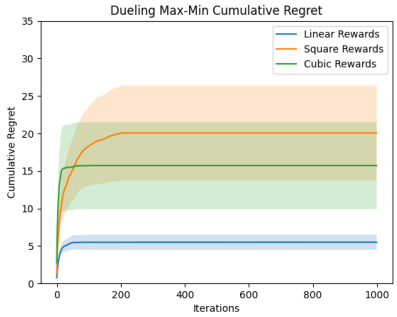


(a) Varying K. (b) Varying N.

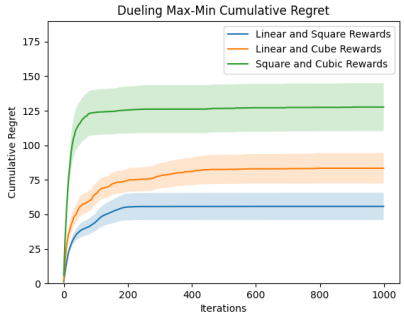
Figure 4: Cumulative regrets of our Dueling ULCB algorithm with varying K and varying N.

Following this, we run experiments for the case where we have base arms $1, 2, 3, 4$ and the number of agents is 2. The base arm rewards are again $N(\mu_i, 1)$, with $\mu_i = i$, like in the last experiment. We use three different reward functions and their combinations for the purpose of our simulations which are $r(\boldsymbol{\mu}; S) = \sum_{i \in S} \mu_i$, $r(\boldsymbol{\mu}; S) = \sum_{i \in S} \mu_i^3$ and $r(\boldsymbol{\mu}; S) = \sum_{i \in S} (\mu_i \vee 0)^2$ respectively. In Figure 5a we present the cumulative regret of the Dueling Max-Min ULCB algorithm with same rewards for all agents. We vary the reward function between the three reward functions as stated above. In Figure 5b we present the cumulative regret of the Dueling Max-Min ULCB algorithm where the agents share different rewards. We vary the reward function between the three reward functions as stated above. As shown in Figure 5, among all the settings of the reward function, our Dueling Max-Min ULCB Algorithm achieves a clear

sub-linear pattern.



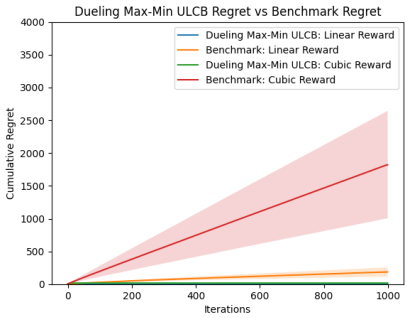
(a) Same reward function for both agents.



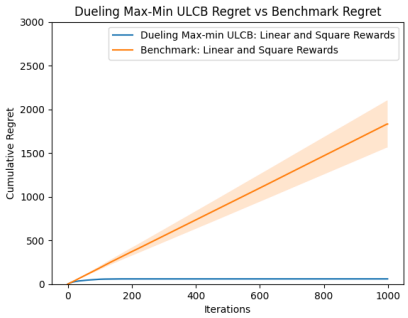
(b) Different reward functions for both agents.

Figure 5: Cumulative regrets of the Dueling Max-Min ULCB algorithm with varying reward functions for both agents.

Finally we compare the Dueling Max-Min ULCB algorithm to a benchmark algorithm which excludes the LCB step from the algorithm and only solves the Max-Min problem using the UCB values. In Figure 6a we do this comparison with the agents sharing the same reward and in Figure 6b, we perform the same experiments with the agents having different rewards. Across all scenarios, the benchmark method exhibits a distinct linear regret, suggesting that relying solely on UCB values is inadequate for resolving our max-min allocation problem. Conversely, our Dueling Max-Min ULCB algorithm demonstrates a significant enhancement in regret.



(a) Same reward function for both agents.



(b) Different reward functions for both agents.

Figure 6: Cumulative regrets of Dueling Max-Min ULCB vs Benchmark with varying reward functions for both agents.

7 Conclusion

Our paper proposes a framework for learning with limited feedback. The constraint on feedback arises from practical considerations, as obtaining feedback, while providing crucial information for online decision-making, can incur significant costs. We apply this framework to various allocation problems, aiming to achieve fairness or stability. Our results uncover important structures indicating that not all information about the allocation is necessary to achieve the desired outcome. Our algorithm has practical applications in real-world scenarios, including online dating markets, job matching, and food allocation.

Future work involves addressing problems with more complex objectives and constraints, where it is possible that limiting feedback to one or a small constant number of agents may not be sufficient to achieve sub-linear regret. An immediate example of this is a constrained optimization problem or an LP with noisy

coefficients. Investigating the trade off between the regret and the size of feedback is interesting and can provide insights into the fundamental structure of the problem. Moreover, our paper currently focuses on the sample complexity of the considered problem and does not discuss the computational issues of the offline oracles. It is an interesting to study the trade-off of computational and sample complexity in practical polynomial-time approximate algorithms.

References

[1] Martin Aleksandrov, Haris Aziz, Serge Gaspers, and Toby Walsh. Online fair division: Analysing a food bank problem. *arXiv preprint arXiv:1502.07571*, 2015.

[2] Martin Aleksandrov and Toby Walsh. Expected outcomes and manipulations in online fair division. In *KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40*, pages 29–43. Springer, 2017.

[3] Martin Aleksandrov and Toby Walsh. Most competitive mechanisms in online fair division. In *KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40*, pages 44–57. Springer, 2017.

[4] Martin Aleksandrov and Toby Walsh. Pure nash equilibria in online fair division. In *IJCAI*, pages 42–48, 2017.

[5] Martin Aleksandrov and Toby Walsh. Strategy-proofness, envy-freeness and pareto efficiency in online fair division with additive utilities. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I 16*, pages 527–541. Springer, 2019.

[6] Martin Aleksandrov and Toby Walsh. Online fair division: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13557–13562, 2020.

[7] Siddhartha Banerjee, Vasilis Gkatzelis, Safwan Hossain, Billy Jin, Evi Micha, and Nisarg Shah. Proportionally fair online allocation of public goods with predictions. *arXiv preprint arXiv:2209.15305*, 2022.

[8] Siddhartha Banerjee, Chamsi Hssaine, and Sean R Sinclair. Online fair allocation of perishable resources. *ACM SIGMETRICS Performance Evaluation Review*, 51(1):55–56, 2023.

[9] Gerdus Benadè, Daniel Halpern, and Alexandros Psomas. Dynamic fair division with partial information. *Advances in neural information processing systems*, 35:3703–3715, 2022.

[10] Gerdus Benadè, Aleksandr M Kazachkov, Ariel D Procaccia, Alexandros Psomas, and David Zeng. Fair and efficient online allocations. *Operations Research*, 2023.

[11] Gerdus Benade, Aleksandr M Kazachkov, Ariel D Procaccia, and Christos-Alexandros Psomas. How to make envy vanish over time. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 593–610, 2018.

[12] Ilai Bistritz, Tavor Baharav, Amir Leshem, and Nicholas Bambos. My fair bandit: Distributed learning of max-min fairness with multi-player bandits. In *International Conference on Machine Learning*, pages 930–940. PMLR, 2020.

[13] Wei Cao, Jian Li, Yufei Tao, and Zhize Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[14] Sarah H Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In *International Conference on Artificial Intelligence and Statistics*, pages 8938–8968. PMLR, 2022.

[15] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[16] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

[17] Giannis Fikioris, Siddhartha Banerjee, and Éva Tardos. Online resource sharing via dynamic max-min fairness: Efficiency, robustness and non-stationarity. *arXiv preprint arXiv:2310.08881*, 2023.

[18] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[19] Evrard Garcelon, Vashist Avadhanula, Alessandro Lazaric, and Matteo Pirotta. Top k ranking for multi-armed bandit with noisy evaluations. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6242–6269. PMLR, 28–30 Mar 2022.

[20] Daniel Golovin. *Max-min fair allocation of indivisible goods.* School of Computer Science, Carnegie Mellon University, 2005.

[21] Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 1057–1066. PMLR, 2018.

[22] Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael I Jordan, and Jacob Steinhardt. Learning equilibria in matching markets with bandit feedback. *Journal of the ACM*, 70(3):1–46, 2023.

[23] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[24] Yasushi Kawase and Hanna Sumita. Online max-min fair allocation. In *International Symposium on Algorithmic Game Theory*, pages 526–543. Springer, 2022.

[25] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[26] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[27] Amir Leshem. Fair multi-agent bandits. *arXiv preprint arXiv:2306.04498*, 2024.

[28] Yuantong Li, Guang Cheng, and Xiaowu Dai. Double matching under complementary preferences. *arXiv preprint arXiv:2301.10230*, 2023.

[29] Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*, 2023.

[30] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

[31] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

[32] Mustafa Lokhandwala and Hua Cai. Dynamic ride sharing using traditional taxis and shared autonomous taxis: A case study of nyc. *Transportation Research Part C: Emerging Technologies*, 97:45–60, 2018.

[33] Evangelos Markakis and Christos-Alexandros Psomas. On worst-case allocations in the presence of indivisible goods. In *International Workshop on Internet and Network Economics*, pages 278–289. Springer, 2011.

[34] Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *Advances in Neural Information Processing Systems*, 35:19956–19970, 2022.

[35] Deepan Muthirayan, Chinmay Maheshwari, Pramod Khargonekar, and Shankar Sastry. Competing bandits in time varying matching markets. In *Learning for Dynamics and Control Conference*, pages 1020–1031. PMLR, 2023.

[36] Thanh Nguyen and Rakesh Vohra. Near-feasible stable matchings with couples. *American Economic Review*, 108(11):3154–3169, 2018.

[37] Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of political Economy*, 92(6):991–1016, 1984.

[38] Alvin E Roth. Deferred acceptance algorithms: History, theory, practice, and open questions. *international Journal of game Theory*, 36(3):537–569, 2008.

[39] Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018.

[40] Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Hongtu Zhu, and Rui Song. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics*, 17(4):2701–2722, 2023.

[41] Hakuei Yamada, Junpei Komiyama, Kenshi Abe, and Atsushi Iwasaki. Learning fair division from bandit feedback. *arXiv preprint arXiv:2311.09068*, 2023.

[42] Mengyan Zhang and Cheng Soon Ong. Quantile bandits for best arms identification. In *International Conference on Machine Learning*, pages 12513–12523. PMLR, 2021.

[43] Ruida Zhou and Chao Tian. Approximate top-m arm identification with heterogeneous reward variances. In *International Conference on Artificial Intelligence and Statistics*, pages 7483–7504. PMLR, 2022.

Appendices
"Active Learning for Fair and Stable Online Allocations"

In this file, we provide all detailed proofs in Sections A-C. In Section A, we provide proofs for the unit demand case which is presented in Section 2 in the main body. We also present some key results in this Section which shall be used through the rest of the Appendix.

A Proofs for Section 2

We further present another Lemma which is fundamental for our main results.

Lemma A.1. *Consider a MAB with N arms, where the i-th arm has noisy rewards $X_i(t) = \mu_i + \epsilon_i(t)$ with $\epsilon_i(t) \sim \sigma^2$-subgaussian, $i = 1, \ldots, N$. Further, denote by $\bar{\nu}_i(t), \underline{\nu}_i(t)$ the UCB and LCB of the reward estimate for the i-th arm at time $t \in \{1, 2, \cdots, T\}$ defined in (2.3) and (2.4). Then for any $\hat{\Delta} \geq 2$, $\delta > 0$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) > C\, \log T)\right\}\right] \leq \frac{2\,N}{T^{\hat{\Delta}^2/2-2}}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\underline{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) > C\, \log T)\right\}\right] \leq \frac{2\,N}{T^{\hat{\Delta}^2/2-2}}$$

where $C = \left(\sqrt{2\alpha} + \hat{\Delta}\right)^2 \sigma^2\, \delta^{-2}$.

Proof of Lemma A.1. We begin by noting that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) > C\, \log T)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (T_i(t-1) > C\, \log T)\right\}\right].
\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\, i\in\mathcal{N}:(|\bar{\nu}_i(t)-\mu_i|\geq\delta)\cap(I_t=i)\cap(T_i(t-1)>C\,\log T)\right\}\right]\\
&\leq\sum_{i\in\mathcal{N}}\sum_{t=1}^{T}\mathbb{P}\left\{(|\bar{\nu}_i(t)-\mu_i|\geq\delta)\cap(T_i(t-1)>C\,\log T)\right\}\\
&\leq\sum_{i\in\mathcal{N}}\sum_{t=1}^{T}\sum_{j=[C\log T]+1}^{T}\mathbb{P}\left[(|\bar{\nu}_i(t)-\mu_i|\geq\delta)\cap(T_i(t-1)=j)\right]
\end{aligned}
$$

where the second step follows from the union bound and the third step follows from the values that $T_i(t-1)$ can take. Further, $[x]$ is the floor function denoting the greatest integer less than x. Now, for any $j=[C\log T]+1,\cdots,T$, we have

$$
\begin{aligned}
&\mathbb{P}\left\{(\bar{\nu}_i(t)-\mu_i\geq\delta)\cap(T_i(t-1)=j)\right\}\\
&\overset{(1)}{\leq}\mathbb{P}\left\{\left(\frac{\sqrt{j}}{\sigma}\left(\hat{\mu}_i(t)-\mu_i\right)\geq\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)\cap(T_i(t-1)=j)\right\}\\
&\overset{(2)}{\leq}\exp\left\{-\frac{1}{2}\left(\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)^2\right\}\\
&\overset{(3)}{\leq}\exp\left\{-\frac{1}{2}\left(\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log T}\right)^2\right\}\\
&\overset{(4)}{\leq}\exp\left\{-\frac{\hat{\Delta}^2}{2}\log T\right\}\\
&\leq\frac{1}{T^{\hat{\Delta}^2/2}}.
\end{aligned}
$$

Note that (1) follows using simple algebra and the fact that $j>C\log T$, (2) follows as the arm rewards are subgaussian, (3) follows as the expression is maximized for $t=T$ as the expression inside the square is positive and monotonically decreasing in t with T fixed (since the square term is minimized at $t=T$), and (4) follows by the definition of $\hat{\Delta}$.

Similarly, for the left tail, we have

$$
\begin{aligned}
&\mathbb{P}\left\{(\bar{\nu}_i(t)-\mu_i\leq-\delta)\cap(T_i(t-1)=j)\right\}\\
&\overset{(5)}{\leq}\mathbb{P}\left\{\left(\frac{\sqrt{j}}{\sigma}\left(\hat{\mu}_i(t)-\mu_i\right)\leq-\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)\cap(T_i(t-1)=j)\right\}\\
&\overset{(6)}{\leq}\exp\left\{-\frac{1}{2}\left(\frac{\delta\sqrt{C\,\log T}}{\sigma}+\sqrt{2\alpha\log t}\right)^2\right\}\\
&\overset{(7)}{\leq}\exp\left\{-\frac{1}{2}\left(\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)^2\right\}\\
&\leq\frac{1}{T^{\hat{\Delta}^2/2}},
\end{aligned}
$$

where (5) follows from algebra and the fact $j>C\log T$, (6) follows from the fact that the rewards are subgaussian and (7) follows as $0<\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}<\frac{\delta\sqrt{C\,\log T}}{\sigma}+\sqrt{2\alpha\log t}$ and the monotone property of the exponential function. Therefore, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\, i\in\mathcal{N}:(|\bar{\nu}_i(t)-\mu_i|\geq\delta)\cap(I_t=i)\cap(T_i(t-1)>C\,\log T)\right\}\right]$$
$$\leq\frac{2\,N}{T^{\hat{\Delta}^2/2-2}}.$$

For the LCB values, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\, i\in\mathcal{N}:(|\underline{\nu}_i(t)-\mu_i|\geq\delta)\cap(I_t=i)\cap(T_i(t-1)>C\,\log T)\right\}\right]$$
$$\leq\sum_{i\in\mathcal{N}}\sum_{t=1}^{T}\mathbb{P}\left\{(|\underline{\nu}_i(t)-\mu_i|\geq\delta)\cap(T_i(t-1)>C\,\log T)\right\}$$
$$\leq\sum_{i\in\mathcal{N}}\sum_{t=1}^{T}\sum_{j=[C\log T]+1}^{T}\mathbb{P}\left[(|\underline{\nu}_i(t)-\mu_i|\geq\delta)\cap(T_i(t-1)=j)\right].$$

And we further have,

$$\mathbb{P}\left\{(\underline{\nu}_i(t)-\mu_i\leq-\delta)\cap(T_i(t-1)=j)\right\}$$
$$\overset{(8)}{\leq}\mathbb{P}\left\{\left(\frac{\sqrt{j}}{\sigma}\left(\mu_i-\hat{\mu}_i(t)\right)\geq\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)\cap(T_i(t-1)=j)\right\}$$
$$\overset{(9)}{\leq}\exp\left\{-\frac{1}{2}\left(\frac{\delta\sqrt{C\,\log T}}{\sigma}-\sqrt{2\alpha\log t}\right)^2\right\}$$
$$\leq\frac{1}{T^{\hat{\Delta}^2/2}},$$

where (8) follows from algebraic manipulation and $j>C\log T$ and (9) follows from the subgaussian arm errors. For the right tail, we have

$$\mathbb{P}\left\{(\underline{\nu}_i(t)-\mu_i\geq\delta)\cap(T_i(t-1)=j)\right\}$$
$$\overset{(10)}{\leq}\mathbb{P}\left\{\left(\frac{\sqrt{j}}{\sigma}\left(\hat{\mu}_i(t)-\mu_i\right)\geq\frac{\delta\sqrt{C\,\log T}}{\sigma}+\sqrt{2\alpha\log t}\right)\cap(T_i(t-1)=j)\right\}$$
$$\leq\frac{1}{T^{\hat{\Delta}^2/2}},$$

where (10) follows from subgaussian arm rewards and the rest of the argument is identical to the argument for left tail of the UCB.

Therefore

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\, i\in\mathcal{N}:(|\underline{\nu}_i(t)-\mu_i|\geq\delta)\cap(I_t=i)\cap(T_i(t-1)>C\,\log T)\right\}\right]$$
$$\leq\frac{2\,N}{T^{\hat{\Delta}^2/2-2}}.$$

Hence we are done. □

Corollary 1. *Under the conditions of Lemma A.1, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\, i\in\mathcal{N}:(|\bar{\nu}_i(t)-\mu_i|\geq\delta)\cap(I_t=i)\right\}\right]\leq N\,C\log T+\frac{2\,N}{T^{\hat{\Delta}^2/2-2}}$$

where $C = \left(\sqrt{2\alpha} + \hat{\Delta}\right)^2 \sigma^2\, \delta^{-2}$.

Proof of Corollary 1. Note that for the event in question we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) > C\log T)\right\}\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) \leq C\log T)\right\}\right].
\end{aligned}$$

Using Lemma A.1 we see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) > C\log T)\right\}\right] \\
&\leq \frac{2N}{T^{\hat{\Delta}^2/2-2}}.
\end{aligned}$$

Further, we see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (|\bar{\nu}_i(t) - \mu_i| \geq \delta) \cap (I_t = i) \cap (T_i(t-1) \leq C\log T)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : (I_t = i) \cap (T_i(t-1) \leq C\log T)\right\}\right] \\
&\leq N\, C\, \log T.
\end{aligned}$$

Hence we are done. □

The following result is present in [14]; we furnish this in our work for completeness.

Lemma A.2. *Consider a MAB with N arms having arm rewards $X_i(t) = \mu_i + \epsilon_i(t)$ with $\epsilon_i(t)$, σ^2-subgaussian and μ_i as the true arm reward. Further, denote by $\bar{\nu}_i(t), \underline{\nu}_i(t)$ the UCB and LCB of the reward estimate for the i-th arm at time $t \in \{1, 2, \cdots, T\}$ defined in* (2.3) *and* (2.4)*. Then*

$$\mathbb{P}\left(\bar{\nu}_i(t) \leq \mu_i\right) \leq \frac{1}{t^{\alpha-1}} \quad \textit{and} \quad \mathbb{P}\left(\underline{\nu}_i(t) \geq \mu_i\right) \leq \frac{1}{t^{\alpha-1}}.$$

Proof of Lemma A.2. The proof follows by decomposing the event in terms of the number of times arm i has been pulled. Note that we may pull arm i at most t times if we are at time instance t.

$$
\begin{aligned}
&\mathbb{P}\left(\bar{\nu}_i(t) \leq \mu_i\right) \\
&= \sum_{\tau=1}^{t} \mathbb{P}\left(\bar{\nu}_i(t) \leq \mu_i \mid T_i(t-1) = \tau\right) \mathbb{P}\left(T_i(t-1) = \tau\right) \\
&\leq \sum_{\tau=1}^{t} \mathbb{P}\left(\bar{\nu}_i(t) \leq \mu_i \mid T_i(t-1) = \tau\right) \\
&\leq \sum_{\tau=1}^{t} \mathbb{P}\left(\frac{\sqrt{\tau}\left(\mu_i - \hat{\mu}_i(t)\right)}{\sigma} \geq \sqrt{2\log t^{\alpha}}\right) \\
&\leq \sum_{\tau=1}^{t} \frac{1}{t^{\alpha}} \\
&= \frac{1}{t^{\alpha-1}}.
\end{aligned}
$$

The proof for $\mathbb{P}\left(\underline{\nu}_i(t) \geq \mu_i\right)$ is similar. Hence we are done. □

Proof of Theorem 1. To establish this result we will first show a concentration bound for the UCB and the LCB estimates of the arms. Subsequently, we will establish that if the arms are sufficiently pulled then our algorithm chooses the correct max-min allocation. Namely, we show two steps-given an allocation not exploring the true minimum in the allocation occurs at most $O(\log T)$ times and given any two allocations, not choosing the allocation with higher minimal reward occurs at most $O(\log T)$ times. Finally we combine the results which results in our regret bound.

First, we establish that for any $\phi \in \mathcal{M}$,

$$
\mathbb{P}\left(\exists\, \phi \in \mathcal{M} : \min_{j \in \mathcal{K}} \bar{\nu}^{j}_{\phi(j)}(t) \leq \min_{j \in \mathcal{K}} \mu^{j}_{\phi(j)}\right) \leq \frac{K\,N}{t^{\alpha-1}}.
$$

To observe this, note that

$$
\begin{aligned}
&\mathbb{P}\left(\exists\, \phi \in \mathcal{M} : \min_{j \in \mathcal{K}} \bar{\nu}^{j}_{\phi(j)}(t) \leq \min_{j \in \mathcal{K}} \mu^{j}_{\phi(j)}\right) \\
&\overset{\text{(I)}}{\leq} \mathbb{P}\left(\exists\, j_1 \in \mathcal{K}, i_1 \in \mathcal{N} : \bar{\nu}^{j_1}_{i_1}(t) \leq \mu^{j_1}_{i_1}\right) \\
&\overset{\text{(II)}}{\leq} \sum_{j_1=1}^{K} \sum_{i=1}^{N} \mathbb{P}\left(\bar{\nu}^{j_1}_{i_1}(t) \leq \mu^{j_1}_{i_1}\right) \\
&\leq \frac{K\,N}{t^{\alpha-1}}.
\end{aligned}
$$

where (I) follows as there exits some $j_1 \in \mathcal{K}$ and $i_1 = \phi(j_1)$ such that

$$
\bar{\nu}^{j_1}_{i_1}(t) = \min_{j \in \mathcal{K}} \bar{\nu}^{j}_{\phi(j)}(t) \leq \min_{j \in \mathcal{K}} \mu^{j}_{\phi(j)} \leq \mu^{j_1}_{i_1}.
$$

(II) follows from probability laws and where the last line follows from Lemma A.2. Note that we may apply the same lemma here as this is indeed a MAB with each agent-item pair forming an arm. Next we establish that the number of discordant pairs between the UCB arm estimates and the true rewards till time T is $O(\log T)$. Namely, we control the event $E_1^{\phi}(t)$ defined as

$$
\left\{\exists\, \phi_1 \in \mathcal{M}\backslash\Phi^*,\, \phi_2 \in \Phi^*,\, j_1 \in \mathcal{K} : (I_t = (j_1, \phi_1(j_1))) \cap \left(\min_{j \in \mathcal{K}} \bar{\nu}^{j}_{\phi_1(j)}(t) \geq \min_{j \in \mathcal{K}} \bar{\nu}^{j}_{\phi_2(j)}(t)\right)\right\}
$$

which represents a mismatch of ordering between the truth and the estimated. Further define the events

$$A_1^{\phi}(t) = \left\{ min_{j\in\mathcal{K}}\, \bar{\nu}^{j}_{\phi_2(j)}(t) > \min_{j\in\mathcal{K}} \mu^{j}_{\phi_2(j)} \right\} \text{ and } B_1^{\phi}(t) = \left\{ I_t = \phi_1(j_1),\, j_1 \in \arg\min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)} \right\}$$

which respectively represent the events that the minimal UCB estimate is greater than the minimal true reward and the arm to explore belongs to the true minima of the chosen allocation. Therefore

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t) \cap A_1^{\phi}(t) \cap B_1^{\phi}(t)\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t) \cap B_1^{\phi}(t)^c\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{A_1^{\phi}(t)^c\right\}\right].
\end{aligned}$$

Note that on the event $E_1^{\phi}(t) \cap A_1^{\phi}(t) \cap B_1^{\phi}(t)$, one has

$$\begin{aligned}
&\bar{\nu}^{j_1}_{\phi_1(j_1)}(t) - \mu^{j_1}_{\phi_1(j_1)} \\
&\overset{(1)}{\geq} \min_{j\in\mathcal{K}} \bar{\nu}^{j}_{\phi_1(j)}(t) - \min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)} \\
&\overset{(2)}{\geq} \min_{j\in\mathcal{K}} \bar{\nu}^{j}_{\phi_2(j)}(t) - \min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)} \\
&\overset{(3)}{\geq} \min_{j\in\mathcal{K}} \mu^{j}_{\phi_2(j)} - \min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)} \\
&\overset{(4)}{\geq} \Delta_{\min},
\end{aligned}$$

where $(1), (2)$ follows from the event $E_1^{\phi}(t)$ as $\min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)} = \mu^{j_1}_{\phi_1(j_1)}$, (3) follows from the event $A_1^{\phi}(t)$ and (4) follows from Assumption 1 due to the event $B_1^{\phi}(t)$. On the event $E_1^{\phi}(t) \cap B_1^{\phi}(t)^c$, consider any $j_2 \in \arg\min_{j\in\mathcal{K}} \mu^{j}_{\phi_1(j)}$. In this case, we note that one of the two events hold-either $\mu_{\text{maxmin}} - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2$ or $\mu^{j_1}_{\phi_1(j_1)} - \mu^{j_2}_{\phi_1(j_2)} \geq \Delta_{\min}/2$. This is because we know that $\mu_{\text{maxmin}} - \mu^{j_2}_{\phi_1(j_2)} > \Delta_{\min}$ and then we divide the event based on which of the former two quantities $\mu^{j_1}_{\phi_1(j_1)}$ is closer to. Hence

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t) \cap B_1^{\phi}(t)^c\right\}\right] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t) \cap \left(\mu_{\text{maxmin}} - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2\right)\right\}\right]}_{(a)} \\
+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{E_1^{\phi}(t) \cap \left(\mu^{j_1}_{\phi_1(j_1)} - \mu^{j_2}_{\phi_1(j_2)} \geq \Delta_{\min}/2\right)\right\}\right]}_{(b)}.
\end{aligned}$$

Further note that for the event in (a), $\exists\, j_3 \in \mathcal{K}$ such that

$$\bar{\nu}^{j_3}_{\phi_2(j_3)}(t) = \min_{j\in\mathcal{K}} \bar{\nu}^{j}_{\phi_2(j)}(t) \leq \min_{j\in\mathcal{K}} \bar{\nu}^{j}_{\phi_1(j)}(t) = \bar{\nu}^{j_1}_{\phi_1(j_1)}(t);$$

however, since $\phi_2 \in \Phi^*$,

$$\mu^{j_3}_{\phi_2(j_3)} \geq \mu_{\text{maxmin}} > \mu^{j_1}_{\phi_1(j_1)} + \Delta_{\min}/2.$$

Consider the event $A_2^\phi(t) = \{\forall\, j \in \mathcal{K} : \bar{\nu}^j_{\phi_2(j)}(t) \geq \mu^j_{\phi(j)}\}$. Note that on

$$E_1^\phi(t) \cap \left(\mu_{\text{maxmin}} - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2\right) \cap A_2^\phi(t)$$

one has

$$\begin{aligned}
&\bar{\nu}^{j_1}_{\phi_1(j_1)}(t) - \mu^{j_1}_{\phi_1(j_1)} \\
&\overset{(4)}{\geq} \bar{\nu}^{j_3}_{\phi_2(j_3)}(t) - \mu^{j_1}_{\phi_1(j_1)} \\
&\overset{(5)}{\geq} \mu^{j_3}_{\phi_2(j_3)}(t) - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2
\end{aligned}$$

where $(4), (5)$ hold due to the event in question. Therefore

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{E_1^\phi(t) \cap \left(\mu_{\text{maxmin}} - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2\right)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{E_1^\phi(t) \cap \left(\mu_{\text{maxmin}} - \mu^{j_1}_{\phi_1(j_1)} \geq \Delta_{\min}/2\right) \cap A_2^\phi(t)\right\}\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{A_2^\phi(t)^c\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{I_t = \phi_1(j_1),\, \bar{\nu}^{j_1}_{\phi_1(j_1)}(t) - \mu^{j_1}_{\phi_1(j_1)} > \Delta_{\min}/2\right\}\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{A_2^\phi(t)^c\right\}\right].
\end{aligned}$$

On the event for (b) note that both $\underline{\nu}^{j_1}_{\phi_1(j_1)}(t) \leq \underline{\nu}^{j_2}_{\phi_1(j_2)}(t)$ and $\mu^{j_1}_{\phi_1(j_1)} > \mu^{j_2}_{\phi_1(j_2)} + \Delta_{\min}/2$ since the arm chosen to explore has minimal reward with respect to the LCB estimates but the true rewards have reverse ordering. Essentially the chosen arm is not the true minimal arm. Consider the event

$$A_3^\phi(t) = \left\{\exists\, \phi_2,\, j \in \mathcal{K} : \underline{\nu}^{j_2}_{\phi_1(j_2)}(t) < \mu^{j_2}_{\phi(j_2)}\right\}$$

Therefore, on the event for

$$E_1^\phi(t) \cap \left(\mu^{j_1}_{\phi_1(j_1)} - \mu^{j_2}_{\phi_1(j_2)} \geq \Delta_{\min}/2\right) \cap A_3^\phi(t)$$

, we have

$$\begin{aligned}
&\mu^{j_1}_{\phi_1(j_1)} - \underline{\nu}^{j_1}_{\phi_1(j_1)}(t) \\
&\geq \mu^{j_1}_{\phi_1(j_1)} - \underline{\nu}^{j_2}_{\phi_1(j_2)}(t) \\
&\mu^{j_1}_{\phi_1(j_1)} - \mu^{j_2}_{\phi_1(j_2)} \geq \Delta_{\min}/2
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{E_1^\phi(t) \cap \left(\mu^{j_1}_{\phi_1(j_1)} - \mu^{j_2}_{\phi_1(j_2)} \geq \Delta_{\min}/2\right)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{I_t = \phi_1(j_1),\, \mu^{j_1}_{\phi_1(j_1)} - \underline{\nu}^{j_1}_{\phi_1(j_1)}(t) > \Delta_{\min}/2\right\}\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{A_3^\phi(t)^c\right\}\right]
\end{aligned}$$

Combining these, we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{E_1^{\phi}(t)\right\}\right] \\
&\leq \underbrace{\mathbb{E}\left[\mathbf{1}\left\{\exists\, j_1\in\mathcal{K},\, \phi_1\in\mathcal{M}:\, (I_t=\phi_1(j_1))\cap\left(\bar{\nu}^{j_1}_{\phi_1(j_1)}(t)-\mu^{j_1}_{\phi_1(j_1)}>\Delta_{\min}\right)\right\}\right]}_{\text{I}} \\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{I_t=\phi_1(j_1),\, \bar{\nu}^{j_1}_{\phi_1(j_1)}(t)-\mu^{j_1}_{\phi_1(j_1)}>\Delta_{\min}/2\right\}\right]}_{\text{II}}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{I_t=\phi_1(j_1),\, \mu^{j_1}_{\phi_1(j_1)}-\underline{\nu}^{j_1}_{\phi_1(j_1)}(t)>\Delta_{\min}/2\right\}\right]}_{\text{III}} \\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{A_1^{\phi}(t)^c\right\}\right]}_{\text{IV}}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{A_2^{\phi}(t)^c\right\}\right]}_{\text{V}}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{A_3^{\phi}(t)^c\right\}\right]}_{\text{VI}}.
\end{aligned}
$$

Using the argument at the beginning, we have term IV less than

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,\phi_2\in\mathcal{M}:\ \min_{j\in\mathcal{K}}\bar{\nu}^{j}_{\phi_2(j)}(t)\leq\min_{j\in\mathcal{K}}\mu^{j}_{\phi_2(j)}\right\}\right] \\
&\leq\sum_{t=1}^{T}\mathbb{P}\left\{\exists\,\phi_2\in\mathcal{M}:\ \min_{j\in\mathcal{K}}\bar{\nu}^{j}_{\phi_2(j)}(t)\leq\min_{j\in\mathcal{K}}\mu^{j}_{\phi_2(j)}\right\} \\
&\leq N\,K\left(1+\sum_{t=2}^{\infty}\frac{1}{t^{\alpha-1}}\right) \\
&\leq N\,K\left(1+\int_{1}^{\infty}\frac{1}{t^{\alpha-1}}dt\right) \\
&\leq\frac{N\,K(\alpha-1)}{\alpha-2}.
\end{aligned}
$$

Using a similar argument for V and VI, we have the same bound for each of these terms. For terms I-III, we use Corollary 1 to get

$$
\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{E_1^{\phi}(t)\right\}\right]\leq\frac{3\,K\,N\,(\alpha-1)}{\alpha-2}+3\,K\,N\left(C\,\log T+\frac{2}{T^{\hat{\Delta}^2/2-2}}\right).
$$

Therefore, noting that

$$
R_T\leq\Delta_{\max}\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{E_1^{\phi}(t)\right\}\right]
$$

we get our final bound as

$$
3\,\Delta_{\max}\left[\frac{K\,N\,(\alpha-1)}{\alpha-2}+\,K\,N\left(C\,\log T+\frac{2}{T^{\hat{\Delta}^2/2-2}}\right)\right].
$$

□

We may obtain a coarser regret bound where there is no identifiability condition for the max-min solutions. Denote by $\Phi(\delta)=\{\phi\in\mathcal{M}\backslash\Phi^*:\ \max_{\phi\in\Phi^*}\min_{j\in\mathcal{K}}\mu^{j}_{\phi(j)}-\min_{j\in\mathcal{K}}\mu^{j}_{\phi(j)}\leq\delta\}$ the set of all allocations which are at most δ away from the optimal allocation.

PROPOSITION A.1. *Let* $\left|\mu_{i_1}^{j_1} - \mu_{i_2}^{j_2}\right| \leq \Delta_{max}$ *for all* $i_1, i_2 \in \mathcal{N}$, $j_1, j_2 \in \mathcal{K}$ *and some* $\Delta_{\max} > 0$ *hold. Then regret of Algorithm 1 satisfies*

$$R_T \leq 6\,\Delta_{\max}\,N\,K\left[\frac{\alpha-1}{\alpha-2}+1\right] + 4.5^{2/3}\left(\Delta_{\max}\,N\,K\,\left(\sqrt{2\alpha}+2\right)^2\sigma^2\log T\right)^{1/3}T^{2/3}.$$

Proof of Proposition A.1. Define, for any $\delta > 0$,

$$\Phi(\delta) = \left\{\phi:\ \mu_{\text{maxmin}} - \min_{j\in\mathcal{K}}\mu_{\phi(j)}^{j} > \delta\right\}.$$

We know that the regret for Algorithm 1 is given as

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{\text{maxmin}} - \mu_{\phi_t(j)}^{j}\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{\text{maxmin}} - \mu_{\phi_t(j)}^{j}\right)\mathbf{1}\left\{\phi_t \in \Phi(\Delta_T)\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{\text{maxmin}} - \mu_{\phi_t(j)}^{j}\right)\mathbf{1}\left\{\phi_t \in \Phi(\Delta_T)^c\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{\text{maxmin}} - \mu_{\phi_t(j)}^{j}\right)\mathbf{1}\left\{\phi_t \in \Phi(\Delta_T)\right\}\right] + \Delta_T\,T.
\end{aligned}$$

Note that the first term allows us to use Theorem 1 with gap Δ_T. Therefore, we have

$$R_T \leq 3\,\Delta_{\max}\,N\,K\left[\frac{\left(\sqrt{2\alpha}+2\right)^2\,\sigma^2}{\Delta_T^2}\,\log T + 2\,\frac{\alpha-1}{\alpha-2}+2\right] + \Delta_T\,T.$$

Optimizing over Δ_T, we get the optimum value of Δ_T to be

$$\Delta_T = \left(6\,\Delta_{\max}\,N\,K\,\left(\sqrt{2\alpha}+2\right)^2\sigma^2\,T^{-1}\,\log T\right)^{1/3}.$$

Substituting, the value, we get,

$$R_T \leq 6\,\Delta_{\max}\,N\,K\left[\frac{\alpha-1}{\alpha-2}+1\right] + 4.5^{2/3}\left(\Delta_{\max}\,N\,K\,\left(\sqrt{2\alpha}+2\right)^2\sigma^2\log T\right)^{1/3}T^{2/3}$$

and hence we are done. □

Note that Proposition A.1 exhibits a rate $O((\log T\,T^2)^{1/3})$ for the regret. This is different from the $O(\sqrt{T\,\log T})$ rate for the regret in the classical MAB setting. This is because the analysis we perform provides is a rate which is of the order $\Delta_{\min}^{-2}$ instead of $\Delta_{\min}^{-1}$. The key reason for this is that the regret analysis is extremely challenging if we decompose it in terms of the allocations.

Proof of Proposition 2.1.

Define for all i

$$\mathcal{G}_i^K = \{\text{ the set of all subsets of size } K \text{ with } \mu_i \text{ as true minimum reward}\}.$$

Note that for any i with $\mu_i > \mu_{(K)}$, we have $\mathcal{G}_i = \emptyset$. Then regret can be written as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t \in \mathcal{G}_i^K\right\}\right].$$

Note that

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t=i\right\}\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t\neq i\right\}\right]}_{\text{(II)}}.$$

Now, for (I), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t=i\right\}\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{\exists\ j\in G^*:\ I_t=i,\ \bar{\nu}_i(t)\geq\bar{\nu}_j(t),\ \mu_i<\mu_j\right\}\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\sum_{j\in G^*}\Delta_i(K)\mathbf{1}\left\{\exists\ j\in G^*:\ I_t=i,\ \bar{\nu}_i(t)\geq\bar{\nu}_j(t),\ \mu_i<\mu_j\right\}\right]
\end{aligned}$$

since we pull arm i, which not in the true optimal set and there exists some $j\in G^*$ which was not selected in the G_t. By [14, Lemma C.4.] and noting that $\Delta(i,j)\geq\Delta_i(K)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t=i\right\}\right]\leq K\sum_{i\in G(0)}\left(\frac{8\,\sigma^2\alpha\,\log T}{\Delta_i(K)}+\frac{\Delta_i(K)\,\alpha}{\alpha-2}\right).$$

For (II), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t\neq i\right\}\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\sum_{j:\,\mu_j\geq\mu_{(K)}}\Delta_i(K)\mathbf{1}\left\{I_t=j,\ \underline{\nu}_i(t)\geq\underline{\nu}_j(t),\ \mu_i<\mu_j\right\}\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\sum_{j:\,\mu_i<\mu_j<\mu_{(K)}}\sum_{l\in G^*}\Delta_i(K)\mathbf{1}\left\{I_t=j,\ \bar{\nu}_l(t)\leq\bar{\nu}_j(t),\ \mu_l>\mu_j\right\}\right]
\end{aligned}$$

where the first term on the right hand side follows from the event that some incorrect arm is chosen which belongs to the optimal set G^* and the second term follows from the event that the arm selected lies between the K-th arm and the i-th arm. In this case, note that this arm was selected in the set G_t instead of some arm l in the set G^*. Therefore again, using [14, Lemma C.4.], we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t\neq i\right\}\right]\\
&\leq\sum_{i\in G(0)}\sum_{j:\,\mu_j\geq\mu_{(K)}}\Delta_i(K)\,\frac{8\,\sigma^2\,\alpha\log T}{\Delta^2(i,j)}+\frac{\Delta_i(K)\,\alpha}{\alpha-2}\\
&+\sum_{i\in G(0)}\sum_{j:\,\mu_i<\mu_j<\mu_{(K)}}\sum_{l\in G^*}\Delta_i(K)\,\frac{8\,\sigma^2\,\alpha\log T}{\Delta^2(l,j)}+\frac{\Delta_i(K)\,\alpha}{\alpha-2}\\
&\leq\sum_{i\in G(0)}\frac{8\,K\,\sigma^2\,\alpha\log T}{\Delta_i(K)}+\sum_{i\in G(0)}\frac{8\,(N-K)\,K\,\sigma^2\,\alpha\log T}{\Delta_i(K)}+\sum_{i\in G(0)}\left(\frac{K\,\Delta_i(K)\,\alpha}{\alpha-2}+\frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2}\right).
\end{aligned}
$$

This implies that

$$
R_T\leq\sum_{i\in G(0)}\left(2\,K+(N-K)\,K\right)\frac{8\,\sigma^2\alpha\,\log T}{\Delta_i(K)}+\sum_{i\in G(0)}\left(\frac{2\,K\,\Delta_i(K)\,\alpha}{\alpha-2}+\frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2}\right)
$$

Now, we establish the gap-independent regret bound. Using this regret decomposition with respect to $\mathcal{G}_i^K$ and noting that $G(0)\cap G(\Delta_T)=G(\Delta_T)$, we have

$$
R_T\leq\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(\Delta_T)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right]+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)\cap G(\Delta_T)^c}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right].
$$

The second term is equivalent to

$$
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)\cap G(\Delta_T)^c}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right]=\sum_{i\in G(0)\cap G(\Delta_T)^c}\Delta_i(K)\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right]
$$

and hence the second term is less than $\Delta_T\,T$ since each $\Delta_i(K)\leq\Delta_T$ and

$$
\sum_{i\in G(0)\cap G(\Delta_T)^c}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{G_t\in\mathcal{G}_i^K\right\}\right]\leq T.
$$

Using the fact that on $G(\Delta_T)$, $\Delta_i(K)>\Delta_T$ we know that the first term is less than

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in G(0)}\Delta_i(K)\mathbf{1}\left\{G_t\in\mathcal{G}_i^K,\ I_t\neq i\right\}\right]\\
&\leq\frac{8\,(N-K)\,K\,\sigma^2\alpha\,\log T}{\Delta_T}+K\sum_{i\in G(0)}\frac{\Delta_i(K)\,\alpha}{\alpha-2}\\
&+\frac{8\,(N-K)\,K\,\sigma^2\,\alpha\log T}{\Delta_T}+\frac{8\,(N-K)^2\,K\,\sigma^2\,\alpha\log T}{\Delta_T}+\sum_{i\in G(0)}\left(\frac{K\,\Delta_i(K)\,\alpha}{\alpha-2}+\frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2}\right).
\end{aligned}
$$

This implies that

$$
R_T\leq T\,\Delta_T+\left(2\,(N-K)\,K+(N-K)^2\,K\right)\frac{8\,\sigma^2\alpha\,\log T}{\Delta_T}+\sum_{i\in G(0)}\left(\frac{2\,K\,\Delta_i(K)\,\alpha}{\alpha-2}+\frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2}\right)
$$

Choosing

$$\Delta_T = \sqrt{\frac{((N-K+1)^2-1)\ K\, 8\,\sigma^2\alpha\ \log T}{T}}$$

gives us

$$\begin{aligned} R_T \leq\ & 2\sqrt{((N-K+1)^2-1)\ K\, 8\,\sigma^2\alpha\, T\ \log T} \\ & + \sum_{i\in G(0)} \left(\frac{2\,K\,\Delta_i(K)\,\alpha}{\alpha-2} + \frac{(N-K)\,K\,\Delta_i(K)\,\alpha}{\alpha-2} \right). \end{aligned}$$

Hence we are done.

□

B Proofs for Section 3

B.1 Proofs for Max-Min Allocation

We establish our main results by solving a more general problem. Consider a CMAB with N base arms having true rewards $\mu_1, \mu_2, \cdots, \mu_N$. Denote

$$\mathcal{A}^m \subseteq \left\{ S :\ S \in 2^{\mathcal{N}}\ , |S| \leq m \right\} \tag{B.1}$$

as the set of arm combinations with capacity constraint that at most m arms are included. Note that in Sections 2 and 3, the parameter m is 1 and N, respectively. Further, note that in Section 3, $\bigcup_{j=1}^{K} \mathcal{A}_j = \mathcal{A}^m$. In our setting, for an element in $\mathcal{A}^m$, the reward is given by a function defined as $r : \mathcal{K} \times \mathbb{R}^N \times \mathcal{A}^m \to \mathbb{R}$. Hence we have a separate reward function for each agent. Define $\boldsymbol{\mu}(S) = \sum_{i\in S} \mu_i$. Our objective is to find

$$\arg\max_{\phi\in\bar{\mathcal{M}}} \min_{1\leq j\leq K} r^j(\boldsymbol{\mu}; \phi(j))$$

where we recall that

$$\mathcal{M} = \{\phi : \mathcal{K} \to \mathcal{A}^m,\ \phi(i) \neq \phi(j)\}$$

is the set of all ordered sets of size K from $2^{\mathcal{A}^m}$, which has each element as $[\phi(1), \phi(2), \cdots, \phi(K)]$ with $\phi(i)$ being the set of goods/super-arm assigned to agent i. Note that the $\phi(i)$'s need not be disjoint which is the case when we have allocations. As in Section 3, define the max-min as $\phi^* = \mathcal{O}_1(\boldsymbol{\mu}, \mathcal{A}^m, K)$ and $j^* = \mathcal{O}_2(\boldsymbol{\mu}, \phi^*, K)$. We subsequently present our main results which can be used to establish the results of Section 3.

Lemma B.1. *For a fixed allocation $\phi \in \mathcal{M}$, we have*

$$\mathbb{P}\left(\min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) \leq \min_{1\leq j\leq K} r^j(\boldsymbol{\mu}; \phi(j)) \right) \leq \frac{N}{t^{\alpha-1}}.$$

Proof of Lemma B.1. Note that the event

$$\left\{ \min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) \leq \min_{1\leq j\leq K} r^j(\boldsymbol{\mu}; \phi(j)) \right\}$$

implies that there exists some i such that

$$r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) = \min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) \leq \min_{1\leq j\leq K} r^j(\boldsymbol{\mu}; \phi(j)) \leq r^i(\boldsymbol{\mu}; \phi(i))$$

which in turn implies $\exists\, i'$ such that $\bar{\nu}_{i'}(t) \leq \mu_{i'}$ by the monotonicity condition of Assumption 2. Therefore

$$\mathbb{P}\left(\min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t);\phi(j)) \leq \min_{1\leq j\leq K} r^j(\boldsymbol{\mu};\phi(j))\right) \leq \mathbb{P}\left(\exists\ i' \text{ such that } \ \bar{\nu}_i(t) \leq \mu_{i'}\right)$$
$$\leq \sum_{i=1}^{N} \mathbb{P}\left(\bar{\nu}_i(t) \leq \mu_{i'}\right) \leq \frac{N}{t^{\alpha-1}},$$

where the last line follows from Lemma A.2. □

Define for each time point t and $\delta > 0$,

$$U_t^{(1)}(\delta) = \big\{\exists\ \phi \in \mathcal{M},\ i,j \in \mathcal{K} : (I_t = \phi(i)) \cap \left(r^i(\bar{\boldsymbol{\nu}}(t);\phi(i)) \geq r^j(\bar{\boldsymbol{\nu}}(t);\phi(j))\right)$$
$$\cap \left(r^i(\boldsymbol{\mu};\phi(i)) + \delta < r^j(\boldsymbol{\mu};\phi(j))\right)\big\} \tag{B.2}$$

as the event which indicates that there is an allocation at time t, for which, the rewards calculated using the UCB values are not correctly ordered with respect to the true rewards; and similarly define

$$L_t^{(1)}(\delta) = \big\{\exists\ \phi \in \mathcal{M},\ i,j \in \mathcal{K} : (I_t = \phi(i)) \cap \left(r^i(\underline{\nu}(t);\phi(i)) \leq r^j(\underline{\nu}(t);\phi(j))\right)$$
$$\cap \left(r^i(\boldsymbol{\mu};\phi(i)) > r^j(\boldsymbol{\mu}) + \delta\right)\big\} \tag{B.3}$$

as the event which indicates that the rewards calculated using the LCB values of the arms are not correctly ordered. We further define

$$M_t = \{\exists\, \phi_1 \in \mathcal{M}\backslash\Phi^*,\ \phi_2 \in \Phi^*,\ i \in \mathcal{K} :$$
$$(I_t = \phi_1(i)) \cap \left(\min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t);\phi_1(j)) \geq \min_{1\leq j\leq K} r^j(\bar{\boldsymbol{\nu}}(t);\phi_2(j))\right)$$
$$\cap \left(\min_{1\leq j\leq K} r^j(\boldsymbol{\mu};\phi_1(j)) < \min_{1\leq j\leq K} r^j(\boldsymbol{\mu};\phi_2(j))\right)\Big\}$$

as the set of allocations with mismatched minimum assignments.

We are now ready to state our main results. Our first result shows that the existence of an allocation containing an incorrect ordering using UCB or LCB estimates on the reward function may happen at most $O(\log T)$ times.

Theorem 6. *Let I_t denote the super arm chosen at time t by Algorithm 2. Under Assumptions 2- 3,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(U_t^{(1)}(\delta))\right] \leq N\left[\frac{\left(\sqrt{2\alpha}+2\right)^2 c^2\, m^2\, \sigma^2}{\delta^2}\ \log T + \frac{\alpha-1}{\alpha-2} + 2\right]$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(L_t^{(1)}(\delta))\right] \leq N\left[\frac{\left(\sqrt{2\alpha}+2\right)^2 c^2\, m^2\, \sigma^2}{\delta^2}\ \log T + \frac{\alpha-1}{\alpha-2} + 2\right].$$

Proof of Theorem 6. The main idea of the proof is to reduce the problem to an MAB. To do this we show that on an event of high probability there exists some explored arm which has either not been sufficiently pulled or falls into a region of low probability when the event in question is true. We start with observing that for any $1 \leq t \leq T$, we have

$$U_t^{(1)}(\delta) = \left(U_t^{(1)}(\delta) \cap \left\{\left(r^j(\bar{\boldsymbol{\nu}}(t);\phi(j)) > r^j(\boldsymbol{\mu};\phi(j))\right) \cap \left(r^i(\bar{\boldsymbol{\nu}}(t);\phi(i)) > r^i(\boldsymbol{\mu};\phi(i))\right)\right\}\right)$$
$$\bigcup \left(U_t^{(1)}(\delta) \cap \left\{\left(r^j(\bar{\boldsymbol{\nu}}(t);\phi(j)) \leq r^j(\boldsymbol{\mu};\phi(j))\right) \cup \left(r^i(\bar{\boldsymbol{\nu}}(t);\phi(i)) \leq r^i(\boldsymbol{\mu};\phi(i))\right)\right\}\right).$$

Now, on

$$U_t^{(1)}(\delta) \cap \left\{ \left(r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) > r^j(\boldsymbol{\mu}; \phi(j)) \right) \cap \left(r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) > r^i(\boldsymbol{\mu}; \phi(i)) \right) \right\},$$

there exists $i' \in \phi(i)$, such that

$$\begin{aligned}
|\bar{\nu}_{i'}(t) - \mu_{i'}| &\overset{(1)}{\geq} \frac{1}{|\phi(i)|} \sum_{i' \in \phi(i)} |\bar{\nu}_{i'}(t) - \mu_{i'}| \\
&\overset{(2)}{\geq} \frac{1}{m} \sum_{i' \in \phi(i)} |\bar{\nu}_{i'}(t) - \mu_{i'}| \\
&\overset{(3)}{\geq} \frac{1}{c\,m} \left| r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) - r^i(\boldsymbol{\mu}; \phi(i)) \right| \\
&\overset{(4)}{\geq} \frac{r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) - r^i(\boldsymbol{\mu}; \phi(i))}{c\,m} \\
&\overset{(5)}{\geq} \frac{r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\boldsymbol{\mu}\phi(i))}{c\,m} \\
&\overset{(6)}{\geq} \frac{r^j(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i))}{c\,m} \\
&\overset{(7)}{\geq} \frac{\delta}{c\,m}.
\end{aligned}$$

(1) holds as there exits one value greater than the average. (2) holds as $|\phi(i)| \leq m$. (3) holds due to Assumption 2. (4)-(6) hold due to the event considered. (7) holds due to Assumption 3. Note that

$$\left(r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) \leq r^i(\boldsymbol{\mu}; \phi(i)) \right) \cup \left(r^j(\bar{\boldsymbol{\nu}}(t); \phi(j)) \leq r^j(\boldsymbol{\mu}; \phi(j)) \right) \subseteq \left\{ \exists\, i' \in \mathcal{N} : \bar{\nu}_{i'}(t) \leq \mu_{i'} \right\}$$

by Assumption 2. Thus, we have

$$\begin{aligned}
U_t^{(1)}(\delta) \subseteq &\left\{ \exists\, i' \in \mathcal{N} : \left(|\bar{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\delta}{c\,m} \right) \cap \left(I_t = i' \right) \right\} \\
&\bigcup \left\{ \exists\, i' \in \mathcal{N} : \bar{\nu}_{i'}(t) \leq \mu_{i'} \right\}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ U_t^{(1)}(\delta) \right\} \right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \exists\, i' \in \mathcal{N} : \left(|\bar{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\delta}{c\,m} \right) \cap \left(I_t = i' \right) \right\} \right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \exists\, i' \in \mathcal{N} : \bar{\nu}_{i'}(t) \leq \mu_{i'} \right\} \right].
\end{aligned}$$

For the first term, using Corollary 1, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \exists\, i' \in \mathcal{N} : \left(|\bar{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\delta}{c\,m} \right) \cap \left(I_t = i' \right) \right\} \right] \leq N\,C\,\log T + \frac{2\,N}{T^{\hat{\Delta}^2/2-2}}$$

where $C = \left(\sqrt{2\alpha} + \hat{\Delta}\right)^2 m^2\, c^2 \sigma^2\, \delta^{-2}$ with $\hat{\Delta} \geq 2$. For the third term, using Lemma B.1, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i' \in \mathcal{N} : \bar{\nu}_{i'}(t) \leq \mu_{i'}\right\}\right] &\leq \sum_{i' \in \mathcal{N}} \sum_{t=1}^{T} \mathbb{P}\left(\bar{\nu}_{i'}(t) \leq \mu_{i'}\right) \\ &\leq N\left(1 + \sum_{t=2}^{\infty} \frac{1}{t^{\alpha-1}}\right) \\ &\leq N\,\frac{\alpha-1}{\alpha-2}.\end{aligned}$$

Combining all bounds and taking $\hat{\Delta} = 2$, the first result follows. The second result can be similarly derived.

We observe that for any $1 \leq t \leq T$, we have

$$\begin{aligned}L_t^{(1)}(\delta) = &\left(L_t^{(1)}(\delta) \cap \left\{\left(r^j(\underline{\boldsymbol{\nu}}(t);\phi(j)) < r^j(\boldsymbol{\mu};\phi(j))\right) \cap \left(r^i(\underline{\boldsymbol{\nu}}(t);\phi(i)) < r^i(\boldsymbol{\mu};\phi(i))\right)\right\}\right) \\ &\bigcup \left(L_t^{(1)}(\delta) \cap \left\{\left(\left(r^j(\underline{\boldsymbol{\nu}}(t);\phi(j)) \geq r^j(\boldsymbol{\mu};\phi(j))\right) \cup \left(r^i(\underline{\boldsymbol{\nu}}(t);\phi(i)) \geq r^i(\boldsymbol{\mu};\phi(i))\right)\right)\right\}\right).\end{aligned}$$

On

$$L_t^{(1)}(\delta) \cap \left\{\left(r^j(\underline{\boldsymbol{\nu}}(t);\phi(j)) < r^j(\boldsymbol{\mu};\phi(j))\right) \cap \left(r^i(\underline{\boldsymbol{\nu}}(t);\phi(i)) < r^i(\boldsymbol{\mu};\phi(i))\right)\right\}$$

there exists an $i' \in \phi(i)$ such that

$$\begin{aligned}|\mu_{i'} - \underline{\nu}_{i'}(t)| &\overset{(1)}{\geq} \frac{1}{|\phi(i)|} \sum_{i' \in \phi(i)} |\mu_{i'} - \underline{\nu}_{i'}(t)| \\ &\overset{(2)}{\geq} \frac{1}{c\,m} \left|r^i(\boldsymbol{\mu};\phi(i)) - r^i(\underline{\boldsymbol{\nu}}(t);\phi(i))\right| \\ &\overset{(3)}{\geq} \frac{1}{c\,m} \left(r^i(\boldsymbol{\mu};\phi(i)) - r^i(\underline{\boldsymbol{\nu}}(t);\phi(i))\right) \\ &\overset{(4)}{\geq} \frac{1}{c\,m} \left(r^i(\boldsymbol{\mu};\phi(i)) - r^j(\underline{\boldsymbol{\nu}}(t);\phi(j))\right) \\ &\overset{(5)}{\geq} \frac{1}{c\,m} \left(r^i(\boldsymbol{\mu};\phi(i)) - r^j(\boldsymbol{\mu};\phi(j))\right) \\ &\overset{(6)}{\geq} \frac{\delta}{c\,m}.\end{aligned}$$

Again, (1) occurs by property of the mean, (2) by Assumption 2 and (3)-(5) are due to the set considered and (6) is due to Assumption 3. Also, note that,

$$\left(r^j(\underline{\boldsymbol{\nu}}(t);\phi(j)) \geq r^j(\boldsymbol{\mu};\phi(j))\right) \cup \left(r^i(\underline{\boldsymbol{\nu}}(t);\phi(i)) \geq r^i(\boldsymbol{\mu};\phi(i))\right) \subseteq \left\{\exists\, i' \in \mathcal{N} : \mu_{i'} \leq \underline{\nu}_{i'}(t)\right\}.$$

Therefore

$$\begin{aligned}L_t^{(1)}(\delta) \subseteq &\left\{\exists\, i' \in \mathcal{N} : \left(|\underline{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\delta}{c\,m}\right) \cap \left(I_t = i'\right)\right\} \\ &\bigcup \left\{\exists\, i' \in \mathcal{N} : \underline{\nu}_{i'}(t) \geq \mu_{i'}\right\}.\end{aligned}$$

The rest of the proof is identical to the previous part. □

Using Theorem 6, we obtain our next result.

THEOREM 7. *Let I_t denote the super arm chosen at time t using Algorithm 2. Then under Assumptions 2-3, one has*

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}(M_t)\right] \leq 2\,N\left[\frac{\left(\sqrt{2\alpha}+2\right)^2 c^2\, m^2\, \sigma^2}{\tilde{\Delta}_{\min}^2}\,\log T + \frac{\alpha-1}{\alpha-2} + 2\right]$$

Proof of Theorem 7. Again, the idea is similar to Theorem 6 where we reduce the problem to an MAB setting where we show that the explored arm is either not sufficiently pulled or the event falls in a region of low probability. We start by defining some events which we shall use throughout the proof. Define

$$A_{\phi}^{(2)}(t) = \left(r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) > r^i(\boldsymbol{\mu}; \phi_1(i))\right)$$

as the event that the UCB estimate for the reward is higher than the true reward and

$$E_{\phi}^{(2)}(t) = \left\{ i \in \arg\min_{1 \le j \le K} r^j(\boldsymbol{\mu}; \phi_1(j)) \right\}.$$

Note that on the event

$$\begin{aligned} M_t = \{ \exists\, \phi_1 \in \mathcal{M} \backslash \Phi^*,\, \phi_2 \in \Phi^*,\, i \in \mathcal{K} : \\ (I_t = \phi(i)) \cap \left(\min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_1(j)) \ge \min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_2(j)) \right) \\ \cap \left(\min_{1 \le j \le K} r^j(\boldsymbol{\mu}; \phi_1(j)) < \min_{1 \le j \le K} r^j(\boldsymbol{\mu}; \phi_2(j)) \right) \Big\} \end{aligned}$$

we have

$$r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) \ge \min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_1(j)) \ge \min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_2(j)).$$

Further,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(M_t\right)\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(M_t \cap A_{\phi}^{(2)}(t)^c\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(M_t \cap A_{\phi}^{(2)}(t) \cap E_{\phi}^{(2)}(t)\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(M_t \cap A_{\phi}^{(2)}(t) \cap E_{\phi}^{(2)}(t)^c\right)\right]. \end{aligned}$$

On $M_t \cap A_{\phi}^{(2)}(t) \cap E_{\phi}^{(2)}(t)$, one has

$$\begin{aligned} |\bar{\nu}_{i'}(t) - \mu_{i'}| &\ge \frac{1}{m} \sum_{j \in \phi_1(i)} |\bar{\nu}_j(t) - \mu_j| \\ &\ge \frac{1}{m\,c} \left| r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right| \end{aligned}$$

where the first inequality follows from the property of mean with $|\phi(i)| \le m$ and the second inequality follows from Assumption 2. Thus,

$$\begin{aligned} |\bar{\nu}_{i'}(t) - \mu_{i'}| &\ge \frac{1}{m\,c} \left(r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right) \\ &\ge \frac{1}{m\,c} \left(\min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_1(j)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right) \\ &\ge \frac{1}{m\,c} \left(\min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_2(j)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right) \\ &\ge \frac{1}{m\,c} \left(\min_{1 \le j \le K} r^j(\bar{\boldsymbol{\nu}}(t); \phi_2(j)) - \min_{1 \le j \le K} r^j(\boldsymbol{\mu}; \phi_1(j)) \right) \quad \ge \frac{\tilde{\Delta}_{\min}}{m\,c}. \end{aligned}$$

where the lines three and four follow by the set considered and the last inequality follows from Assumption 3.

Note that the probability of the set

$$\begin{aligned}
&\mathbb{P}\left\{M_t \cap A_\phi^{(2)}(t)^c\right\} \\
&\leq \mathbb{P}\left(r^i(\bar{\boldsymbol{\nu}}(t);\phi_1(i)) \leq r^i(\boldsymbol{\mu};\phi_1(i))\right) \\
&\leq \mathbb{P}\left(\exists\, i' \in \mathcal{N} :\ \bar{\nu}_{i'}(t) \leq \mu_{i'}\right) \\
&\leq \frac{N}{t^{\alpha-1}}
\end{aligned}$$

where the last two lines follow from Assumption 2 and Lemma A.2 respectively. On the event, $M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c$, define, without loss of generality, $j' = \arg\min_{1\leq j\leq K} r^j(\boldsymbol{\mu};\phi_1(j))$ (again we may take any $j' \in \arg\min_{1\leq j\leq K} r^j(\boldsymbol{\mu};\phi_1(j))$ and the proof does not change). Note that on $M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c$, one of the following events occur- either $r^{j'}(\boldsymbol{\mu};\phi_1(j')) - \min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_1(j)) > \tilde{\Delta}_{\min}/2$ or $\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j)) - r^{j'}(\boldsymbol{\mu};\phi(j')) > \tilde{\Delta}_{\min}/2$. Therefore on the event

$$M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c \cap \left(\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j)) - r^{j'}(\boldsymbol{\mu};\phi(j')) > \tilde{\Delta}_{\min}/2\right)$$

we have

$$\begin{aligned}
&\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j)) > r^{j'}(\boldsymbol{\mu};\phi_1(j')) + \tilde{\Delta}_{\min}/2 \\
&\min_{j\in\mathcal{K}} r^j(\bar{\boldsymbol{\nu}}(t);\phi_2(j)) \leq \min_{j\in\mathcal{K}} r^j(\bar{\boldsymbol{\nu}}(t);\phi_1(j)) = r^{j'}(\bar{\boldsymbol{\nu}}(t);\phi_1(j')).
\end{aligned}$$

Consider the event $A_\phi^{(3)}(t) = \{\min_{j\in\mathcal{K}} r^j(\bar{\boldsymbol{\nu}}(t);\phi_2(j)) \geq \min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j))\}$ Therefore on the event

$$M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c \cap \left(\min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j)) - r^{j'}(\boldsymbol{\mu};\phi(j')) > \tilde{\Delta}_{\min}/2\right) \cap A_\phi^{(3)}(t),$$

we have

$$\begin{aligned}
&r^{j'}(\bar{\boldsymbol{\nu}}(t);\phi_1(j')) - r^{j'}(\boldsymbol{\mu};\phi_1(j')) \\
&\geq \min_{j\in\mathcal{K}} r^j(\bar{\boldsymbol{\nu}}(t);\phi_2(j)) - r^{j'}(\boldsymbol{\mu};\phi_1(j')) \\
&\geq \min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_2(j)) - r^{j'}(\boldsymbol{\mu};\phi_1(j')) \geq \tilde{\Delta}_{\min}/2
\end{aligned}$$

Therefore on the event in question, there exists some $i' \in \phi_1(j')$ such that

$$\begin{aligned}
|\bar{\nu}_{i'}(t) - \mu_{i'}| &\geq \frac{1}{m}\sum_{i\in\phi(j')} |\bar{\nu}_i(t) - \mu_i| \\
&\geq \frac{1}{mc}\left|r^{j'}(\bar{\boldsymbol{\nu}}(t);\phi_1(j')) - r^{j'}(\boldsymbol{\mu};\phi_1(j'))\right| \\
&\geq \frac{\tilde{\Delta}_{\min}}{2\,m\,c}.
\end{aligned}$$

Note that the event $A_\phi^{(3)}(t)^c$ has probability at most $N/t^{\alpha-1}$. Finally, on the event

$$M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c \cap \left(r^{j'}(\boldsymbol{\mu};\phi_1(j')) - \min_{j\in\mathcal{K}} r^j(\boldsymbol{\mu};\phi_1(j)) > \tilde{\Delta}_{\min}/2\right),$$

we have

$$\begin{aligned}
&r^i(\boldsymbol{\mu};\phi_1(i)) > r^{j'}(\boldsymbol{\mu};\phi_1(j')) + \tilde{\Delta}_{\min}/2 \\
&r^i(\underline{\boldsymbol{\nu}}(t);\phi_1(i)) \leq r^{j'}(\underline{\boldsymbol{\nu}}(t);\phi_1(j')).
\end{aligned}$$

Therefore the event $M_t \cap A_\phi^{(2)}(t) \cap E_\phi^{(2)}(t)^c$ implies that $L_t^{(1)}(\delta)$, defined in (B.3), occurs for ϕ_1. Hence, combining the bounds, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(M_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(L_t^{(1)}(\delta))\right] + 2\sum_{t=1}^{T}\frac{N}{t^{\alpha-1}} \\ + \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\left(\exists\, i' \in \mathcal{N} :\, |\bar{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\tilde{\Delta}_{\min}}{m\,c}\right) \cap \left(I_t = i'\right)\right\}\right] \\ + \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\left(\exists\, i' \in \mathcal{N} :\, |\bar{\nu}_{i'}(t) - \mu_{i'}| \geq \frac{\tilde{\Delta}_{\min}}{2\,m\,c}\right) \cap \left(I_t = i'\right)\right\}\right].\end{aligned}$$

Therefore the result follows using Theorem 6 and Corollary 1 on the first, the third and the fourth term respectively. □

Proof of Theorem 2. We note that

$$R_t \leq \tilde{\Delta}_{\max}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\text{the correct allocation is not chosen at time } t\right\}\right] \leq \tilde{\Delta}_{\max}\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(M_t)\right]$$

and hence the result follows using Theorem 7. □

We may also provide an instance independent regret bound in this setting

Proposition B.1. *Let for any $\phi \in \mathcal{M}$ and pair $i, j \in \mathcal{K}$, $i \neq j$,*

$$\left|r^j(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i))\right| \leq \tilde{\Delta}_{\max}$$

hold. Then, under Assumption 2, the regret for Algorithm 2 satisfies

$$R_T \leq 3\,\Delta_{\max}\,N\left[\frac{\alpha-1}{\alpha-2}+2\right] + 4.5^{2/3}\left(\Delta_{\max}\,N^3\,c^2\,\left(\sqrt{2\alpha}+2\right)^2\sigma^2\log T\right)^{1/3}T^{2/3}$$

Proof of Proposition B.1. The proof is identical to Proposition A.1. □

B.2 Replenishing Items, Same Rewards

In this section, we investigate a variant of our problem in which resources are continuously replenished, ensuring that each agent receives a unique set of resources. Here, we have N items that are replenished after each agent's turn. During a turn, an agent is assigned a specific set of items, which are replenished before the next agent's turn begins. This process repeats until all agents have been served, marking the completion of one time instance. The system operates under two primary rules: ensuring no two agents receive the same set of items and disclosing only the base reward for one agent to the system, reflecting our commitment to active feedback.

The lack of constraints on the selection of super-arms is a direct consequence of the arms' resampling and the departure from considering partitions exclusively. Therefore the problem reduces to finding

$$\max_{\phi\in\mathcal{M}}\min_{j\in\mathcal{K}} r(\boldsymbol{\mu}; \phi(j)). \tag{B.4}$$

This problem is essentially finding the top K super-arms in the CMAB setting by revealing only one super-arm's rewards. Define any K-th best super-arm as S^* and define the top K super-arm set chosen at time t as $\mathcal{S}_t$. In this case, the cumulative regret reduces to

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(r(\boldsymbol{\mu}; S^*) - \min_{S\in\mathcal{S}_t} r(\boldsymbol{\mu}; S)\right)\right].$$

To solve this problem, we can use Algorithm 2 with a minor change. That is, the first oracle, when given a vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_N)$ and a reward function r, solves the problem (B.4) and the second oracle is simply a sorting oracle. We can now establish the regret bound for Algorithm 2 in this setting.

Corollary 2. *Let Assumptions 2-3 hold. The regret bound for Algorithm 2 satisfies*

$$R_T \leq 3\,\tilde{\Delta}_{\max} N\left[\frac{\left(\sqrt{2\alpha}+2\right)^2 c^2 N^2 \sigma^2}{\tilde{\Delta}_{\min}^2}\,\log T + \frac{\alpha-1}{\alpha-2} + 2\right].$$

Proof of Corollary 2. The proof follows immediately from Theorem 6. □

An interesting case is when the reward is defined as

$$r(\boldsymbol{\mu}; S) = \mathbb{E}\left[f\left(\sum_{i\in S} X_i\right)\right] \tag{B.5}$$

with $f(\cdot)$ being a known function. The main question of interest here is-what conditions on f are sufficient to establish the desired regret bounds.

Proposition B.2. *Let the function $f(\cdot)$ is monotone and L-Lipschitz.. Then, for algorithm 2,*

$$R_T = O(m^2 N\,\log T).$$

Proof of Proposition B.2. With some notation abuse we denote $\sum_{i\in S} X_i = S^T\mathbf{X}$. Note that for any $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we may take $\tilde{\mathbf{X}}(t) = (\mathbf{X}(t) - \boldsymbol{\mu}) + \boldsymbol{\nu}$ such that $r(\boldsymbol{\mu}; S) = \mathbb{E}[f(S^\mathsf{T}\mathbf{X}]$ and $r(\boldsymbol{\nu}; S) = \mathbb{E}[f(S^\mathsf{T}\tilde{\mathbf{X}}(t))]$. Therefore,

$$\begin{aligned}
|r(\boldsymbol{\mu}; S) - r(\boldsymbol{\nu}; S)| &= \left|\mathbb{E}[f(S^\mathsf{T}\mathbf{X}(t))] - \mathbb{E}[f(S^\mathsf{T}\tilde{\mathbf{X}}(t))]\right| \\
&\leq \mathbb{E}\left|f(S^\mathsf{T}\mathbf{X}(t)) - f(S^\mathsf{T}\tilde{\mathbf{X}}(t))\right| \\
&\leq L\,\mathbb{E}\left|S^\mathsf{T}\mathbf{X}(t) - S^\mathsf{T}\tilde{\mathbf{X}}(t)|\right| \\
&= L\,\left|S^\mathsf{T}\left(\boldsymbol{\mu} - \boldsymbol{\nu}\right)\right| \\
&= L\,|\boldsymbol{\mu}(S) - \boldsymbol{\nu}(S)| \\
&\leq L\sum_{i\in S}|\mu_i - \nu_i|\,.
\end{aligned}$$

Also, if $\nu_i \geq \mu_i$, for all $i \in S$, then

$$S^\mathsf{T}\tilde{\mathbf{X}}(t) = S^\mathsf{T}\mathbf{X}(t) + S^\mathsf{T}\left(\mu^{(2)} - \mu^{(1)}\right) \geq S^\mathsf{T}\mathbf{X}(t).$$

This implies $f(S^\mathsf{T}\tilde{\mathbf{X}}(t)) \geq f(S^\mathsf{T}\mathbf{X}(t))$ by the monotone property of f. Therefore

$$r(\boldsymbol{\nu}; S) - r(\boldsymbol{\mu}; S) = \mathbb{E}\left[f(S^\mathsf{T}\tilde{\mathbf{X}}(t)) - f(S^\mathsf{T}\mathbf{X}(t))\right] \geq 0.$$

Hence Assumption 2 is established and thus the proof follows. □

B.3 Proofs for Minimal Envy Allocation

In this setting we shall again work in the general regime where $\mathcal{M} = \{\phi : \phi : \mathcal{K} \to 2^{\mathcal{N}},\ \phi(i) \neq \phi(j)\}$, $r^j : \mathbb{R}^N \times \mathcal{A}_j \to \mathbb{R}$ and $\bigcup_{j=1}^{K} \mathcal{A}_j = \mathcal{A}^m$. Thus we have a capacity constraint on each set with $|A| \leq m$. We establish our results in this regime and then present our results with $m = N$.

Lemma B.2. *For any fixed allocation $\phi \in \mathcal{M}$, one has*

$$\mathbb{P}\left(ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi) > ev(\boldsymbol{\mu}, \phi)\right) \leq \frac{2\,K\,(K-1)\,N}{t^{\alpha-1}}$$

and

$$\mathbb{P}\left(ev(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) < ev(\boldsymbol{\mu}, \phi)\right) \leq \frac{2\,K\,(K-1)\,N}{t^{\alpha-1}}.$$

Proof of Lemma B.2. We start by fixing $i, j \in \mathcal{K}$. Note that in the event

$$\{ev_{i\to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi) > ev_{i\to j}(\boldsymbol{\mu}, \phi)\}$$

one of the following must be true- either

$$r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) > r^i(\boldsymbol{\mu}; \phi(j))$$

or

$$r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) < r^i(\boldsymbol{\mu}; \phi(i)).$$

This is immediate from the fact that on the event in question $ev_{i\to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi) > 0$ since the inequality is strict and hence

$$\left(\left(r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\bar{\boldsymbol{\nu}}(t); \phi(i))\right) > \left(r^i(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i))\right)\right).$$

Note that the statement holds true even if the right hand side is less than 0. This immediately implies

$$\begin{aligned}
&\mathbb{P}\left(ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi) > ev(\boldsymbol{\mu}, \phi)\right) \\
&\leq \mathbb{P}\left(ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi) - ev(\boldsymbol{\mu}, \phi) > 0\right) \\
&\leq \mathbb{P}\left(\max_{i,j\in\mathcal{K}} \left(\max\left(r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)), 0\right) - \max\left(r^i(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i)), 0\right)\right) > 0\right) \\
&\leq K\,(K-1)\,\mathbb{P}\left(\left(\max\left(r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)), 0\right) - \max\left(r^i(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i)), 0\right)\right) > 0\right) \\
&\leq K\,(K-1)\left[\mathbb{P}\left(r^i(\underline{\boldsymbol{\nu}}(t); \phi(j)) > r^i(\boldsymbol{\mu}; \phi(j))\right) + \mathbb{P}\left(r^i(\bar{\boldsymbol{\nu}}(t); \phi(i)) < r^i(\boldsymbol{\mu}; \phi(i))\right)\right] \\
&\leq \frac{2\,K\,(K-1)\,N}{t^{\alpha-1}}
\end{aligned}$$

where the third inequality follows from the definition of maximum and the fourth inequality follows from the fact that the event implies the existence of $i, j \in \mathcal{K}$, $i \neq j$ such that the event in question holds. The second last inequality follows from the discussion at the beginning of the proof and the final inequality follows from Lemma A.2. The proof for the other tail bound is exactly identical. □

Lemma B.3. *For any $x_t, y_t \in \mathbb{R}^N$ with $t = 1, 2$, one has*

$$\begin{aligned}
&\max_{i,j\in\mathcal{K}} \max\left(r^i(x_1; \phi(j)) - r^i(y_1; \phi(i)), 0\right) - \max_{i,j\in\mathcal{K}} \max\left(r^i(x_2; \phi(j)) - r^i(y_2; \phi(i)), 0\right) \\
&\leq \left|r^{i^*}(x_1; \phi(j^*)) - r^{i^*}(x_2; \phi(j^*))\right| + \left|r^{i^*}(y_1; \phi(i^*)) - r^{i^*}(y_2; \phi(i^*))\right|
\end{aligned}$$

for some $i^*, j^* \in \mathcal{K}$.

Proof of Lemma B.3. The proof follows by observing that there exists i^*, j^* such that

$$\begin{aligned}
&\max_{i,j\in\mathcal{K}} \max\left(r^i(x_1;\phi(j)) - r^i(y_1;\phi(i)), 0\right) - \max_{i,j\in\mathcal{K}} \max\left(r^i(x_2;\phi(j)) - r^i(y_2;\phi(i)), 0\right) \\
&\leq \max\left(r^{i^*}(x_1;\phi(j^*)) - r^{i^*}(y_1;\phi(i^*)), 0\right) - \max\left(r^{i^*}(x_2;\phi(j^*)) - r^{i^*}(y_2;\phi(i^*)), 0\right).
\end{aligned}$$

Now note that if $r^{i^*}(x_1;\phi(j^*)) - r^{i^*}(y_1;\phi(i^*)) \leq 0$, the last term is less than equal to 0. Hence the result trivially follows. If $r^{i^*}(x_1;\phi(j^*)) - r^{i^*}(y_1;\phi(i^*)) > 0$, then

$$\begin{aligned}
&\max_{i,j\in\mathcal{K}} \max\left(r^i(x_1;\phi(j)) - r^i(y_1;\phi(i)), 0\right) - \max_{i,j\in\mathcal{K}} \max\left(r^i(x_2;\phi(j)) - r^i(y_2;\phi(i)), 0\right) \\
&\leq \left(r^{i^*}(x_1;\phi(j^*)) - r^{i^*}(y_1;\phi(i^*))\right) - \left(r^{i^*}(x_2;\phi(j^*)) - r^{i^*}(y_2;\phi(i^*))\right) \\
&\leq \left|r^{i^*}(x_1;\phi(j^*)) - r^{i^*}(x_2;\phi(j^*))\right| + \left|r^{i^*}(y_1;\phi(i^*)) - r^{i^*}(y_2;\phi(i^*))\right|.
\end{aligned}$$

Thus we are done. □

Define

$$\begin{aligned}
E_t(\delta) = \big\{&\exists\ i,j,i',j',\ \phi \notin \mathcal{E}^* : (I_t = \phi(i) \cup \phi(j)) \cap \\
&\left((i,j) \in \arg\max_{i,j\in\mathcal{K}} \left(\max\left(r^i(\bar{\boldsymbol{\nu}}(t);\phi(j)) - r^i(\underline{\boldsymbol{\nu}}(t);\phi(i)), 0\right)\right)\right) \\
&\cap \left((i',j') \notin \arg\max_{i',j'\in\mathcal{K}} \left(\max\left(r^{i'}(\boldsymbol{\mu};\phi(j')) - r^{i'}(\boldsymbol{\mu};\phi(i'), 0\right)\right)\right) \\
&\cap \left(ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) > ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi)\right) \\
&\cap \left(ev_{i\to j}(\boldsymbol{\mu}, \phi) + \delta < ev_{i'\to j'}(\boldsymbol{\mu}, \phi)\right)\big\}
\end{aligned} \tag{B.6}$$

as the event that for some allocation $\phi \in \mathcal{M}$ the maximal envy is not chosen by the upper estimate at time t. We show that the total size of this event is controlled.

Proposition B.3. *Under Assumption 2, one has*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta))\right] = O\left(\frac{c^2\, m^2\, N \log T}{\delta^2}\right).$$

Proof of Proposition B.3. We establish the result by reducing this problem to a MAB problem based on the arm explored. The remaining event can be reduced to that of low probability. Define the event

$$A_\phi^{(3)}(t) = \left(ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) \geq ev_{i'\to j'}(\boldsymbol{\mu}, \phi)\right).$$

We note that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta) \cap A_\phi^{(3)}(t))\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(A_\phi^{(3)}(t)^c)\right].$$

Note that on the event $E_t(\delta) \cap A_\phi^{(3)}(t)$,

$$\begin{aligned}
&ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) - ev_{i\to j}(\boldsymbol{\mu}, \phi) \\
&\overset{(1)}{\geq} ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) - ev_{i\to j}(\boldsymbol{\mu}, \phi) \\
&\overset{(2)}{\geq} ev_{i'\to j'}(\boldsymbol{\mu}, \phi) - ev_{i\to j}(\boldsymbol{\mu}, \phi).
\end{aligned}$$

where (1) and (2) follow from the event in question. The last expression is greater than δ. Therefore, it is easy to see that

$$ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) - ev_{i\to j}(\boldsymbol{\mu}, \phi) \geq \delta,$$

which implies

$$\left(r^i(\bar{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\underline{\boldsymbol{\nu}}(t); \phi(i))\right) - \left(r^i(\boldsymbol{\mu}; \phi(j)) - r^i(\boldsymbol{\mu}; \phi(i))\right) \geq \delta.$$

A brief explanation of this is as $\delta > 0$, the first term must be positive and the second term being negative only increases the expression in value. This in turn implies,

$$\begin{aligned} &\left|r^i(\bar{\boldsymbol{\nu}}(t); \phi(j)) - r^i(\boldsymbol{\mu}; \phi(j))\right| + \left|r^i(\boldsymbol{\mu}; \phi(i)) - r^i(\underline{\boldsymbol{\nu}}(t); \phi(i))\right| \geq \delta; \\ &\text{which implies} \qquad c \sum_{l\in\phi(j)} |\bar{\nu}_l(t) - \mu_l| + c \sum_{l\in\phi(i)} |\mu_l - \underline{\nu}_l(t)| \geq \delta \end{aligned}$$

using Assumption 2. Therefore, by the property of average, there exists $l_1 \in \phi(i), l_2 \in \phi(j)$ such that $|\phi(i)|\,|\bar{\nu}_{l_1}(t) - \mu_{l_1}| + |\phi(j)|\left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\delta}{c}$ which implies $|\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\delta}{cm}$ since $|\phi(i)|, |\phi(j)| \leq m$. Hence

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta) \cap A_\phi^{(3)}(t))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ (i,j) \in \mathcal{N}\times\mathcal{N} : (I_t = (i,j)) \cap \left(|\bar{\nu}_i(t) - \mu_i| + \left|\mu_j - \underline{\nu}_j(t)\right| \geq \frac{\delta}{cm}\right)\right\}\right].$$

This can be further bounded by

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ i \in \mathcal{N} : (I_t = i) \cap \left(|\bar{\nu}_i(t) - \mu_i| \geq \frac{\delta}{2\,cm}\right)\right\}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ j \in \mathcal{N} : (I_t = j) \cap \left(\left|\mu_j - \underline{\nu}_j(t)\right| \geq \frac{\delta}{2\,cm}\right)\right\}\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta))\right] \\ &\quad \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t(\delta) \cap A_\phi^{(3)}(t))\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(A_\phi^{(3)}(t)^c)\right] \\ &\quad \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ i \in \mathcal{N} : (I_t = i) \cap \left(|\bar{\nu}_i(t) - \mu_i| \geq \frac{\delta}{2\,cm}\right)\right\}\right]}_{\text{(I)}} \\ &\quad + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ j \in \mathcal{N} : (I_t = j) \cap \left(\left|\mu_j - \underline{\nu}_j(t)\right| \geq \frac{\delta}{2\,cm}\right)\right\}\right]}_{\text{(II)}} \\ &\quad + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) < ev_{i'\to j'}(\boldsymbol{\mu}, \phi)\right)}_{\text{(III)}} \end{aligned}$$

Note for (III), we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi) < ev_{i'\to j'}(\boldsymbol{\mu}, \phi)\right) \\
&= \sum_{t=1}^{T} \mathbb{P}\left(\max\left(r^{i'}(\bar{\boldsymbol{\nu}}(t), \phi(j')) - r^{i'}(\underline{\boldsymbol{\nu}}(t), \phi(i')), 0\right) < \max\left(r^{i'}(\boldsymbol{\mu}, \phi(j')) - r^{i'}(\boldsymbol{\mu}, \phi(i')), 0\right)\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(r^{i'}(\bar{\boldsymbol{\nu}}(t), \phi(j')) < r^{i'}(\boldsymbol{\mu}, \phi(j'))\right) + \sum_{t=1}^{T} \mathbb{P}\left(r^{i'}(\underline{\boldsymbol{\nu}}(t), \phi(i')) > r^{i'}(\boldsymbol{\mu}, \phi(i'))\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(\exists\, i \in \mathcal{N} :\, \bar{\nu}_i(t) \leq \mu_i\right) + \sum_{t=1}^{T} \mathbb{P}\left(\exists\, i \in \mathcal{N} :\, \underline{\nu}_i(t) \geq \mu_i\right) \\
&\leq \frac{2\,N\,(\alpha-1)}{\alpha-2}.
\end{aligned}$$

Using Corollary 1 for (I) and (II), and Lemma B.2 for the last term, we get the final bound as

$$\frac{\left(\sqrt{2\alpha}+2\right)^2\, 8\,c^2\,m^2\,N\log T}{\delta^2} + 4\,N + \frac{2\,N\,(\alpha-1)}{\alpha-2}$$

Hence we are done. Note that when $m = N$, this bound is $O(N^3 \log T \delta^{-2})$. □

Proof of Theorem 3. The key idea in this proof is to leverage Proposition B.3 and reduce the problem to the base arms except in a region of low probability. Define, for any $\phi^* \in \mathcal{E}^*$, the event

$$A_{\phi}^{(4)}(t) = \{ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi^*) \leq ev(\boldsymbol{\mu}, \phi^*)\}$$

and the event

$$\begin{aligned}
\text{Env}_t = \{\exists\ i, j \in \mathcal{K}, \phi_1 \in \mathcal{M}\backslash\mathcal{E}^*,\, \phi_2 \in \mathcal{E}^* : (I_t = \phi_1(i) \cup \phi_1(j))) \\
\cap\, (ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1) \leq ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2)) \cap (ev(\boldsymbol{\mu}, \phi_1) > ev(\boldsymbol{\mu}, \phi_2))\}
\end{aligned}$$

which is the event that the optimal envy allocation is not chosen and explored at time t. Using Lemma B.3, note that,

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \left(ev(\boldsymbol{\mu}, \phi_t) - ev(\boldsymbol{\mu}, \phi^*)\right)\right] \\
&\leq \Delta_{e,\max}\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(\phi_t \neq \phi^*\right)\right] \\
&\leq \Delta_{e,\max}\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\text{Env}_t\right\}\right] \\
&\leq \Delta_{e,\max}\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\text{Env}_t \cap A_{\phi}^{(4)}(t)\right\}\right] + \Delta_{e,\max}\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{A_{\phi}^{(4)}(t)^c\right\}\right].
\end{aligned}$$

Invoking Lemma B.2, we know that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{A_{\phi}^{(4)}(t)^c\right\}\right] \leq \sum_{t=1}^{T} \frac{2\,K(K-1)\,N}{t^{\alpha-1}}.$$

Further, on the event $\text{Env}_t \cap A_\phi^{(4)}(t)$, one has

$$\begin{aligned} & ev(\boldsymbol{\mu}, \phi_1) - ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2) \\ & \overset{(1)}{\geq} ev(\boldsymbol{\mu}, \phi_1) - ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2) \\ & \overset{(2)}{\geq} ev(\boldsymbol{\mu}, \phi_1) - ev(\boldsymbol{\mu}, \phi_2) \\ & \overset{(3)}{\geq} \Delta_{e,\min} > 0 \end{aligned}$$

where (1) and (2) hold because of the event in consideration and (3) folds due to Assumption 4. Note that

$$(i,j) = \underset{i_1, j_1 \in \mathcal{K}}{\arg\max} \left(\max \left(r^{i_1}(\bar{\boldsymbol{\nu}}(t); \phi_1(i_1)) - r^{i_1}(\underline{\boldsymbol{\nu}}(t); \phi_1(j_1)), 0 \right) \right)$$

by property of the first oracle. Define the event

$$\text{Err}_t = \{(i,j) \in \underset{i_1, j_1 \in \mathcal{K}}{\arg\max} \left(\max \left(r^{i_1}(\boldsymbol{\mu}; \phi_2(j_1)) - r^{i_1}(\boldsymbol{\mu}; \phi_2(i_1)), 0 \right) \right)\}.$$

Further, without loss of generality, define $(i', j') = \arg\max_{i_1, j_1 \in \mathcal{K}} \left(\max \left(r^{i_1}(\boldsymbol{\mu}; \phi_2(j_1)) - r^{i_1}(\boldsymbol{\mu}; \phi_2(i_1)), 0 \right) \right)$. Note that it does not actually matter what pair we draw as the envy value is same. Note that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \text{Env}_t \cap A_\phi^{(4)}(t) \right\} \right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t \right\} \right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{ \text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c \right\} \right].$$

On $\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t$, we have

$$\begin{aligned} & ev_{i \to j}(\boldsymbol{\mu}, \phi_1) - ev_{i \to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1) \\ & = ev(\boldsymbol{\mu}, \phi_1) - ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1) \\ & \geq \Delta_{e,\min} > 0. \end{aligned}$$

Again, we see that

$$ev_{i \to j}(\boldsymbol{\mu}, \phi_1) - ev_{i \to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1) \geq \Delta_{e,\min}$$

implies

$$\left(r^i(\boldsymbol{\mu}; \phi_1(j)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right) - \left(r^i(\underline{\boldsymbol{\nu}}(t); \phi_1(j)) - r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) \right) \geq \Delta_{e,\min}$$

as $\Delta_{e,\min} > 0$. This in turn implies,

$$\begin{aligned} & \left| r^i(\bar{\boldsymbol{\nu}}(t); \phi_1(i)) - r^i(\boldsymbol{\mu}; \phi_1(i)) \right| + \left| r^i(\boldsymbol{\mu}; \phi_1(j)) - r^i(\underline{\boldsymbol{\nu}}(t); \phi_1(j)) \right| \geq \Delta_{e,\min}; \\ & \text{which implies} \qquad c \sum_{l \in \phi_1(j)} |\bar{\nu}_l(t) - \mu_l| + c \sum_{l \in \phi_1(i)} |\mu_l - \underline{\nu}_l(t)| \geq \Delta_{e,\min} \end{aligned}$$

using Assumption 2. Therefore, there exists $l_1 \in \phi_1(i)$, $l_2 \in \phi_1(j)$ such that

$$|\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left| \mu_{l_2} - \underline{\nu}_{l_2}(t) \right| \geq \frac{\Delta_{e,\min}}{c\, m}.$$

On $\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c$ we have one of the following events occur-

$$ev_{i \to j}(\mu, \phi_1) - ev(\mu, \phi_2) \geq \frac{\Delta_{e,\min}}{2}$$

or

$$ev(\mu, \phi_1) - ev_{i \to j}(\mu, \phi_1) \geq \frac{\Delta_{e,\min}}{2}$$

Therefore on

$$\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c \cap \left(ev_{i\to j}(\mu;\phi_1) - ev(\mu;\phi_2) \geq \frac{\Delta_{e,\min}}{2} \right) \cap (ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2) \leq ev(\mu,\phi_2))$$

we have

$$\begin{aligned}
&ev_{i\to j}(\mu,\phi_1) - ev_{i\to j}(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1)\\
&ev_{i\to j}(\mu,\phi_1) - ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_1)\\
&ev_{i\to j}(\mu,\phi_1) - ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2)\\
&ev_{i\to j}(\mu,\phi_1) - ev(\mu,\phi_2) \geq \frac{\Delta_{e,\min}}{2}.
\end{aligned}$$

Note that this further implies that there exists there exists $l_1 \in \phi_1(i)$, $l_2 \in \phi_1(j)$ such that

$$|\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\Delta_{e,\min}}{2\,c\,m}.$$

Also note that using Lemma B.2, we know that

$$\mathbb{P}\left(ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2) > ev(\mu,\phi_2)\right) \leq \frac{2\,K(K-1)\,N}{t^{\alpha-1}}.$$

Therefore note that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c \cap \left(ev_{i\to j}(\mu;\phi_1) - ev(\mu;\phi_2) \geq \frac{\Delta_{e,\min}}{2} \right)\right\}\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c \cap \left(ev_{i\to j}(\mu;\phi_1) - ev(\mu;\phi_2) \geq \frac{\Delta_{e,\min}}{2} \right) \cap (ev(\underline{\boldsymbol{\nu}}(t), \bar{\boldsymbol{\nu}}(t), \phi_2) \leq ev(\mu,\phi_2))\right\}\right]\\
&\quad + \sum_{t=1}^T \frac{2\,K(K-1)\,N}{t^{\alpha-1}}\\
&\leq \mathbb{E}\left[\mathbf{1}\left\{\exists\,(l_1,l_2) \in \mathcal{N}\times\mathcal{N} :\ |\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\Delta_{e,\min}}{2\,c\,m}\right\}\right] + \frac{2\,K\,(K-1)\,N\,(\alpha-1)}{\alpha-2}.
\end{aligned}$$

Finally on the set

$$\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c \cap \left(ev(\mu,\phi_1) - ev_{i\to j}(\mu,\phi_1) \geq \frac{\Delta_{e,\min}}{2} \right)$$

we have

$$ev_{i\to j}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi_1) \geq ev_{i'\to j'}(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t), \phi_1)$$

but

$$ev_{i\to j}(\boldsymbol{\mu},\phi_1) + \frac{\Delta_{e,\min}}{2} < ev_{i'\to j'}(\boldsymbol{\mu},\phi_1)$$

since the maximal envy is not selected. This is simply the event $E_t(\Delta_{e,\min}/2)$ as defined in (B.6). Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\left\{\text{Env}_t \cap A_\phi^{(4)}(t) \cap \text{Err}_t^c\right\}\right]\\
&\leq \sum_{t=1}^T \mathbb{E}\left[\mathbf{1}(E_t(\Delta_{e,\min}/2))\right] + \mathbb{E}\left[\mathbf{1}\left\{\exists\,(l_1,l_2) \in \mathcal{N}\times\mathcal{N} :\ |\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\Delta_{e,\min}}{2\,c\,m}\right\}\right]\\
&\quad + \frac{K\,(K-1)\,N\,(\alpha-1)}{\alpha-2}.
\end{aligned}$$

which further implies

$$
\begin{aligned}
R_T \leq \Delta_{e,\max} \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}(E_t)\right] \\
+ \Delta_{e,\max}\, \mathbb{E}\left[\mathbf{1}\left\{\exists\,(l_1,l_2) \in \mathcal{N}\times\mathcal{N} :\, |\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\Delta_{e,\min}}{c\,m}\right\}\right] \\
+ \Delta_{e,\max}\, \mathbb{E}\left[\mathbf{1}\left\{\exists\,(l_1,l_2) \in \mathcal{N}\times\mathcal{N} :\, |\bar{\nu}_{l_1}(t) - \mu_{l_1}| + \left|\mu_{l_2} - \underline{\nu}_{l_2}(t)\right| \geq \frac{\Delta_{e,\min}}{2\,c\,m}\right\}\right] \\
+ \Delta_{e,\max}\, \frac{4\,K\,(K-1)\,N\,(\alpha-1)}{\alpha-2}.
\end{aligned}
$$

Thus using Proposition B.3 for the first term and Corollary 1 for the second term, we obtain the final bound as

$$
\frac{\left(\sqrt{2\alpha}+2\right)^2\, 13\,\Delta_{e,\max}\,c^2\,m^2\,N\log T}{\Delta_{e,\min}^2} + 6\,\Delta_{e,\max}\,N + \frac{6\,\Delta_{e,\max}\,K\,(K-1)\,N\,(\alpha-1)}{\alpha-2}.
$$

Hence the proof is concluded. □

Using Theorem 3, we may also obtain an instance independent regret bound for Algorithm 3 based on whether the envy of the chosen allocation is close to the optimal envy or not.

PROPOSITION B.4. *Let for any ϕ_1, ϕ_2 and any pairs $(i,j) \neq (i',j') \in \mathcal{K}$, there exists $\Delta_{e,\max} > 0$ such that*

$$
\left|ev_{i\to j}(\boldsymbol{\mu}, \phi_1) - ev_{i'\to j'}(\boldsymbol{\mu}; \phi_2)\right| \leq \Delta_{e,\max}
$$

hold. Then, under Assumption 2, the regret for Algorithm 3 satisfies

$$
R_T = O(N\,T^{2/3}(\log T)^{1/2}).
$$

Proof of Proposition B.4. Define

$$
G(\delta) = \left\{\phi \,:\, ev(\boldsymbol{\mu},\phi) - \min_{\phi'\in\mathcal{E}^*} ev(\boldsymbol{\mu},\phi') \geq \delta\right\}.
$$

Therefore note that for the regret of Algorithm 3 as defined in (4.3), we have

$$
R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(ev(\boldsymbol{\mu},\phi_t) - ev(\boldsymbol{\mu},\phi^*)\right)\,\mathbf{1}\left\{\phi_t \in G(\Delta_T)^c\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left(ev(\boldsymbol{\mu},\phi_t) - ev(\boldsymbol{\mu},\phi^*)\right)\,\mathbf{1}\left\{\phi_t \in G(\Delta_T)\right\}\right]
$$

Note that the first term is bounded by $\Delta_T\,T$. Using Theorem 3, we know that

$$
\begin{aligned}
R_T &\leq \Delta_T\,T + \frac{\left(\sqrt{2\alpha}+2\right)^2\, 13\,\Delta_{e,\max}\,c^2\,m^2\,N\log T}{\Delta_T^2} + 6\,\Delta_{e,\max}\,N + \frac{6\,\Delta_{e,\max}\,K\,(K-1)\,N\,(\alpha-1)}{\alpha-2} \\
&\leq \frac{3}{2}\left(\left(\sqrt{2\alpha}+2\right)^2\, 13\,\Delta_{e,\max}\,c^2\,m^2\,N\log T\right)^{1/3} T^{2/3} + 6\,\Delta_{e,\max}\,N + \frac{6\,\Delta_{e,\max}\,K\,(K-1)\,N\,(\alpha-1)}{\alpha-2}
\end{aligned}
$$

where the last step follows via optimizing on Δ_T. Hence we are done. □

C Proofs for Stable Allocations

In this section we prove the main results in Section 5.

Lemma C.1. *Let Assumption 2 hold. Then for any $L \subset \mathcal{K}$ and $\phi \in \mathcal{M}$ with $\phi' \in \mathcal{M}_{\phi,L}$, we have $g^L(\mathbf{x}, \mathbf{y}, \phi \to \phi')$ as monotone increasing in $\mathbf{x}$, monotone decreasing in $\mathbf{y}$ where $\mathbf{x} = (x_1, x_2, \cdots, x_N) \in \mathbb{R}^N$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N) \in \mathbb{R}^N$. Further*

$$\begin{aligned}
&\left|g^L(\mathbf{x}_1, \mathbf{y}_1, \phi \to \phi') - g^L(\mathbf{x}_2, \mathbf{y}_2, \phi \to \phi')\right| \\
&\leq \sum_{j \in L} \left(\left|r^j(\mathbf{x}_1, \phi(j)) - r^j(\mathbf{x}_2, \phi(j))\right| + \left|r^j(\mathbf{y}_1, \phi'(j)) - r^j(\mathbf{y}_2, \phi'(j))\right|\right).
\end{aligned}$$

Proof of Lemma C.1. By definition we know that

$$g^L(\mathbf{x}, \mathbf{y}, \phi \to \phi') = \max_{j \in L} r^j(\mathbf{x}, \phi(j)) - r^j(\mathbf{y}, \phi'(j)).$$

Therefore for any $\mathbf{z}_x \geq \mathbf{x}$ (coordinate-wise) we have $r^j(\mathbf{x}, \phi(j)) \leq r^j(\mathbf{z}_x, \phi(j))$ for each $j \in L$. Therefore is is increasing in the first coordinate since maximum is a monotone function. Similarly, for any $\mathbf{z}_y \geq \mathbf{y}$ (coordinate-wise), we have $r^j(\mathbf{y}, \phi(j)) \leq r^j(\mathbf{z}_y, \phi(j))$ for each $j \in L$ which implies it is decreasing in the second coordinate. Finally, for any $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^N$, one has

$$\begin{aligned}
&\left|g^L(\mathbf{x}_1, \mathbf{y}_1, \phi \to \phi') - g^L(\mathbf{x}_2, \mathbf{y}_2, \phi \to \phi')\right| \\
&= \left|\max_{j \in L} \left(r^j(\mathbf{x}_1, \phi(j)) - r^j(\mathbf{y}_1, \phi'(j))\right) - \max_{j \in L} \left(r^j(\mathbf{x}_2, \phi(j)) - r^j(\mathbf{y}_2, \phi'(j))\right)\right| \\
&\leq \left|\max\left[\max_{j \in L} \left(r^j(\mathbf{x}_1, \phi(j)) - r^j(\mathbf{y}_1, \phi'(j)) - r^j(\mathbf{x}_2, \phi(j)) + r^j(\mathbf{y}_2, \phi'(j))\right),\right.\right. \\
&\qquad \left.\left.\max_{j \in L} \left(r^j(\mathbf{x}_2, \phi(j)) - r^j(\mathbf{y}_2, \phi'(j)) - r^j(\mathbf{x}_1, \phi(j)) + r^j(\mathbf{y}_1, \phi'(j))\right)\right]\right| \\
&\leq \max_{j \in L} \left(\left|r^j(\mathbf{x}_1, \phi(j)) - r^j(\mathbf{x}_2, \phi(j))\right| + \left|r^j(\mathbf{y}_1, \phi'(j)) - r^j(\mathbf{y}_2, \phi'(j))\right|\right) \\
&\leq \sum_{j \in L} \left(\left|r^j(\mathbf{x}_1, \phi(j)) - r^j(\mathbf{x}_2, \phi(j))\right| + \left|r^j(\mathbf{y}_1, \phi'(j)) - r^j(\mathbf{y}_2, \phi'(j))\right|\right).
\end{aligned}$$

Hence we are done. □

Lemma C.2. *Under Assumption 5, one has*

$$\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right] \\
&\quad \leq 8\, N\, \kappa^2\, m^2 \left(\sqrt{2\alpha} + \hat{\Delta}\right)^2 c^2\, \sigma^2 \left(\Delta_{\mathcal{M}^*,q}^{-2} + (\eta - \epsilon)^{-2}\right) \log T + \frac{8\, N}{T^{\hat{\Delta}^2/2-2}} + \frac{2\, N\, (\alpha - 1)}{\alpha - 2}.
\end{aligned}$$

Proof. The main idea of the proof is to reduce the problem to an MAB setting except a region of low probability. The key is to divide the problem into multiple events which allow us in reducing each case to an MAB setting.

Note that the event $\{\delta_t^\epsilon = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\}$ implies that for all $\phi \in \mathcal{M}$, there exists $L \subset K$ and $\phi' \in \mathcal{M}|_{\phi,L}$ such that $g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') < \epsilon$. In addition, we have one of the following hold based on the matching returned by Algorithm 4- for all matching in $\mathbb{F}(\boldsymbol{\mu}, \eta)$ there exists $L \subset \mathcal{K}$, $\phi' \in \mathcal{M}|_{\phi,L}$ such

that $g^L(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta$ or there exists a matching $\phi \in \mathbb{F}(\boldsymbol{\mu}, \eta)$ such that for all (L, ϕ') we have $g^L(\bar{\boldsymbol{\nu}}(t), \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta$. That is

$$\begin{aligned}
&\{\delta_t^\epsilon = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\} \\
&\subseteq \left\{\delta_t^\epsilon = 0, \exists\, \phi \in \mathbb{F}(\boldsymbol{\mu}, \eta),\ (L, \phi') \text{ such that } g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta\right\} \\
&\bigcup \{\delta_t^\epsilon = 0, \exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta) \text{ such that} \\
&\qquad\qquad g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta\ \forall\, (L, \phi')\big\}\,.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right] \\
&\leq \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0, \exists\, \phi \in \mathbb{F}(\boldsymbol{\mu}, \eta),\ (L, \phi') \text{ such that } g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta\right\}\right]}_{\text{(I)}} \\
&\quad + \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0, \exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta) \text{ such that } g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta\ \forall\, (L, \phi')\right\}\right]}_{\text{(II)}}.
\end{aligned}$$

We first consider (I) where $\phi \in \mathbb{F}(\boldsymbol{\mu}, \eta)$ is not selected as there exists (L, ϕ') such that $g^L(\bar{\nu}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta$. Note that for any $\phi \in \mathbb{F}(\boldsymbol{\mu}, \eta)$, one has $g^L(\boldsymbol{\mu}; \phi \to \phi') \geq \eta$ for all (L, ϕ'). Hence (I) is bounded by

$$\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0, \exists\, \phi \in \mathbb{F}(\boldsymbol{\mu}, \eta),\ (L, \phi') \text{ such that } g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\ \phi,\ (L, \phi')\ : g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') < \eta\,, g^L(\boldsymbol{\mu}; \phi \to \phi') \geq \eta\right\}\right] \\
&\overset{(1)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} \text{ such that } \mu_i > \bar{\nu}_i(t) \text{ or } \mu_i < \underline{\nu}_i(t)\right\}\right] \\
&\leq \frac{2\,N\,(\alpha - 1)}{\alpha - 2}.
\end{aligned}$$

Here (1) holds by the definition of g^L and the monotonicity of the reward function. An argument for this is as follows- the event in question is equivalent to $\max_{j \in L} r^j(\bar{\boldsymbol{\nu}}(t), \phi(j)) - r^j(\underline{\boldsymbol{\nu}}(t), \phi'(j)) \leq \max_{j \in L} r^j(\boldsymbol{\mu}, \phi(j)) - r^j(\boldsymbol{\mu}, \phi'(j))$ by definition of g^L. This implies there exists a $j \in \mathcal{K}$ such that the maximum on the right side is realized and for that $j \in \mathcal{K}$, we have

$$r^j(\bar{\boldsymbol{\nu}}(t), \phi(j)) - r^j(\underline{\boldsymbol{\nu}}(t), \phi'(j)) \leq r^j(\boldsymbol{\mu}, \phi(j)) - r^j(\boldsymbol{\mu}, \phi'(j)).$$

This further implies that either $r^j(\bar{\boldsymbol{\nu}}(t), \phi(j)) \leq r^j(\boldsymbol{\mu}, \phi(j))$ or $r^j(\underline{\boldsymbol{\nu}}(t), \phi'(j)) > r^j(\boldsymbol{\mu}, \phi'(j))$. Thus by the monotonicity of r^j, we get the inequality. The last step follows from Lemma A.2 and an easy bounding which has been shown in previous proofs.

For (II), we note that as $\phi_t \in \mathcal{M}^* \backslash\ \mathbb{F}(\boldsymbol{\mu}, \eta)$, there exists $(\bar{L}_t, \bar{\phi}'_t)$ such that $g^{\bar{L}_t}(\boldsymbol{\mu}; \phi_t \to \bar{\phi}'_t) < \eta$ but $g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi_t \to \phi') \geq \eta$ for all (L, ϕ'). Further note that the exploration is only the sets $\phi_t(j), \phi'_t(j)$ for

all $j \in L_t$, for some L_t such that $g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') < \epsilon$. In the case where for $(\bar{L}_t, \bar{\phi}'_t) \neq (L_t, \phi'_t)$, that is the exploration is not on the correct active sets, then $g^{L_t}(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi_t \to \phi'_t) < \epsilon$, but $g^{L_t}(\boldsymbol{\mu}; \phi_t \to \phi'_t) \geq \eta$. Hence for the event in (II), one has

$$
\begin{aligned}
&\left\{\delta_t^\epsilon = 0,\, \exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta) \text{ such that } I_t = (L, \phi, \phi'),\right. \\
&\qquad\qquad \left. g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta \;\forall\, (L, \phi')\right\} \\
&\subseteq \left\{\exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta),\, (L, \phi') : I_t = (L, \phi, \phi')\right. \\
&\qquad\qquad \left. g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta\,,\, g^L(\boldsymbol{\mu}; \phi \to \phi') < \eta\right\} \\
&\bigcup \left\{\exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta),\, (L, \phi') : I_t = (L, \phi, \phi')\right. \\
&\qquad\qquad \left. g^L(\boldsymbol{\mu}; \phi_t \to \phi'_t) \geq \eta,\, g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi_t \to \phi'_t) < \epsilon\right\}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^\epsilon = 0,\, \exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta) \text{ such that } I_t = (L, \phi, \phi'),\right.\right. \\
&\qquad\qquad \left.\left. g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta \;\forall\, (L, \phi')\right\}\right] \\
&\leq \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta),\, (L, \phi') : I_t = (L, \phi, \phi')\, g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta\,,\, g^L(\boldsymbol{\mu}; \phi \to \phi') < \eta\right\}\right]}_{\text{(III)}} \\
&+ \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta),\, (L, \phi') : I_t = (L, \phi, \phi')\, g^L(\boldsymbol{\mu}; \phi_t \to \phi'_t) - g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi_t \to \phi'_t) \geq \eta - \epsilon\right\}\right]}_{\text{(IV)}}.
\end{aligned}
$$

For (III), we have

$$
\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta),\, (L, \phi') : I_t = (L, \phi, \phi')\right.\right. \\
&\qquad\qquad \left.\left. g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta\,,\, g^L(\boldsymbol{\mu}; \phi \to \phi') < \eta\right\}\right] \\
&\leq \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, (L, \phi')\,,\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta)\; : I_t = (L, \phi, \phi')\right.\right. \\
&\qquad\qquad g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') - g^L(\boldsymbol{\mu}; \phi \to \phi') \geq \eta - g^L(\boldsymbol{\mu}; \phi \to \phi') \\
&\qquad\qquad \left.\left. \text{and } g^L(\boldsymbol{\mu}; \phi \to \phi') < \eta\right\}\right] \\
&\leq \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, (L, \phi')\,,\, \phi \in \mathcal{M}^* \backslash \mathbb{F}(\boldsymbol{\mu}, \eta)\; : I_t = (L, \phi, \phi')\right.\right. \\
&\qquad\qquad g^L(\bar{\boldsymbol{\nu}}(t); \underline{\boldsymbol{\nu}}(t); \phi \to \phi') - g^L(\boldsymbol{\mu}; \phi \to \phi') \\
&\qquad\qquad \left.\left. \geq \inf_{\mathcal{B}_\phi}\left(\eta - g^L(\boldsymbol{\mu}; \phi \to \phi')\right)\right\}\right].
\end{aligned}
$$

From Assumption 5, we know that $\Delta_{\mathcal{M}^*,q} \leq \inf_{\mathcal{B}_\phi}\left(\eta - g^L(\boldsymbol{\mu}; \phi \to \phi')\right)$ which is positive due. Therefore, using Lemma C.1, one has (III) less than

$$\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\left(L,\phi'\right),\,\phi\in\mathcal{M}^*\backslash\mathbb{F}(\boldsymbol{\mu},\eta):I_t=(L,\phi,\phi')\right.\right.$$
$$\left.\left.\left(\sum_{j\in L}\left|r^j(\bar{\boldsymbol{\nu}}(t);\phi(j))-r^j(\boldsymbol{\mu};\phi(j))\right|+\sum_{j\in L}\left|r^j(\underline{\boldsymbol{\nu}}(t);\phi'(j))-r^j(\boldsymbol{\mu};\phi'(j))\right|\right)\geq\Delta_{\mathcal{M}^*,q}\right\}\right].$$

Using triangle inequality and Assumption 2, this in turn implies that (III) is less than

$$\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\left(L,\phi'\right):I_t=(L,\phi)\,\left(\sum_{j\in L}\sum_{i\in\phi(j)}|\bar{\nu}_i(t)-\mu_i|\right)\geq\frac{\Delta_{\mathcal{M}^*,q}}{2\,c}\right\}\right]$$
$$+\,\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\left(L,\phi'\right):I_t=(L,\phi')\,\left(\sum_{j\in L}\sum_{i\in\phi'(j)}|\underline{\nu}_i(t)-\mu_i|\right)\geq\frac{\Delta_{\mathcal{M}^*,q}}{2\,c}\right\}\right].$$

Using the property that there exists a value greater than the average and that $|L|\leq\kappa$, one has

$$\leq\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,i\in\mathcal{N}:I_t=i,\,|\bar{\nu}_i(t)-\mu_i|\geq\frac{\Delta_{\mathcal{M}^*,q}}{2\,c\,\kappa\,m}\right\}\right]$$
$$+\,\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,i\in\mathcal{N}:I_t=i,\,|\underline{\nu}_i(t)-\mu_i|\geq\frac{\Delta_{\mathcal{M}^*,q}}{2\,c\,\kappa\,m}\right\}\right].$$

From an identical argument, we have (IV) less than

$$\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,i\in\mathcal{N}:I_t=i,\,|\bar{\nu}_j(t)-\mu_j|\geq\frac{(\eta-\epsilon)}{2\,c\,\kappa\,m}\right\}\right]$$
$$+\,\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,i\in\mathcal{N}:I_t=i,\,|\underline{\nu}_i(t)-\mu_i|\geq\frac{(\eta-\epsilon)}{2\,c\,\kappa\,m}\right\}\right]$$

Using Corollary 1 for each term in the final bounds for (III) and (IV), we get

$$\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\delta_t^{\epsilon}=0,\,\exists\,\phi\in\mathcal{M}^*\backslash\mathbb{F}(\boldsymbol{\mu},\eta)\,\text{such that }g^L(\bar{\boldsymbol{\nu}}(t);\underline{\boldsymbol{\nu}}(t);\phi\to\phi')\geq\eta\,\forall\,(L,\phi')\right\}\right]$$
$$\leq 8\,N\,\kappa^2\,m^2\left(\sqrt{2\alpha}+\hat{\Delta}\right)^2\,c^2\,\sigma^2\,\left(\Delta_{\mathcal{M}^*,q}^{-2}+(\eta-\epsilon)^{-2}\right)\log T+\frac{8\,N}{T^{\hat{\Delta}^2/2-2}}.$$

Combining, the terms, the proof is completed. □

Proof of Proposition 5.1. Note that when $\delta_t^{\epsilon}=1$, then at time t there exists $\phi\in\mathcal{M}$ such that $g^L(\underline{\boldsymbol{\nu}}(t);\bar{\boldsymbol{\nu}}(t);\phi\to\phi')\geq\epsilon$ for all L,ϕ'. Also note that since H_0 is true then for all ϕ there exists L,ϕ such that $g^L(\boldsymbol{\mu};\phi\to\phi')<0$. Hence, by the definition of $R_T^{H_0}$, we have

$$R_T^{H_0}\leq\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,\phi,\,(L,\phi'),\,\phi\in\mathcal{M}^*:\right.\right.$$
$$\left.\left.g^L(\bar{\boldsymbol{\nu}}(t);\underline{\boldsymbol{\nu}}(t);\phi\to\phi')\geq\epsilon>0>g^L(\boldsymbol{\mu});\phi\to\phi')\right\}\right]$$
$$\leq\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,j\in\mathcal{K},\,\phi,\phi':r^j(\underline{\boldsymbol{\nu}}(t);\phi(j))-r^j(\bar{\boldsymbol{\nu}}(t);\phi'(j))\geq r^j(\boldsymbol{\mu};\phi(j))-r^j(\boldsymbol{\mu};\phi'(j))\right\}\right].$$

Hence, we can argue that

$$R_T^{H_0} \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\, i \in \mathcal{N} : \underline{\nu}_i(t) \geq \mu_i \text{ or } \bar{\nu}_i(t) \leq \mu_i\right\}\right] \leq \frac{2\,N\,(\alpha-1)}{\alpha-2}.$$

Therefore the proof is completed. □

Proof of Theorem 4. To establish Theorem 4, we divide the problem into two parts based on whether the output belongs to the feasible set or not. This further allows us to reduce the problem to an MAB setting.

Hence we shall have

$$\begin{aligned} R_T^{H_a} &= \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^{\epsilon} = 0\right\}\right] \\ &\leq \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^{\epsilon}(\phi_t) = 0, \phi_t \in \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right]}_{\text{(I)}} \\ &\quad + \underbrace{\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^{\epsilon}(\phi_t) = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right]}_{\text{(II)}}. \end{aligned}$$

From Lemma C.2 we know that for (II), we have

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^{\epsilon}(\phi_t) = 0, \phi_t \notin \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right] \\ &\leq 8\,N\,\kappa^2\,m^2\left(\sqrt{2\alpha} + \hat{\Delta}\right)^2 c^2\,\sigma^2\,\left(\Delta_{\mathcal{M}^*,q}^{-2} + (\eta-\epsilon)^{-2}\right)\log T + \frac{8\,N}{T^{\hat{\Delta}^2/2-2}} + \frac{2\,N\,\alpha}{\alpha-2}. \end{aligned}$$

For (I), we know that there exists (L, ϕ_t') such that $g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi_t \to \phi_t') < \epsilon$ since the null hypothesis is declared true. However, $\phi_t \in \mathbb{F}(\boldsymbol{\mu}, \eta)$.
Therefore $g^L(\boldsymbol{\mu}; \phi_t \to \phi_t') \geq \eta$. Thus,

$$\begin{aligned} &\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\delta_t^{\epsilon}(\phi_t) = 0, \phi_t \in \mathbb{F}(\boldsymbol{\mu}, \eta)\right\}\right] \\ &\leq \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\,(L,\phi') :\, I_t = (L,\phi,\phi'),\, g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') < \epsilon,\, g^L(\boldsymbol{\mu}; \phi \to \phi') \geq \eta\right\}\right] \\ &\leq \mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\,(L,\phi') :\, I_t = (L,\phi,\phi'),\; g^L(\boldsymbol{\mu}; \phi \to \phi') - g^L(\underline{\boldsymbol{\nu}}(t); \bar{\boldsymbol{\nu}}(t); \phi \to \phi') \geq \eta - \epsilon\right\}\right]. \end{aligned}$$

Replicating the same argument as in Lemma C.2, we get (I) to be bounded by

$$\mathbb{E}_{H_a}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\exists\,(L,\phi') \,:\, I_t = (L,\phi,\phi') \right.\right. \\ \left.\left. \left(\sum_{j\in L}\left|r^j(\bar{\boldsymbol{\nu}}(t); \phi'(j)) - r^j(\boldsymbol{\mu}; \phi'(j))\right| + \sum_{j\in L}\left|r^j(\underline{\boldsymbol{\nu}}(t); \phi(j)) - r^j(\boldsymbol{\mu}; \phi(j))\right|\right) \geq \eta - \epsilon\right\}\right].$$

This, as in Lemma C.2, further gives

$$\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\delta_t^{\epsilon}(\phi_t)=0,\phi_t\in\mathbb{F}(\boldsymbol{\mu},\eta)\right\}\right]\\
&\leq \mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,(L,\phi')\,:I_t=(L,\phi,\phi'),\,\left(\sum_{j\in L}\sum_{i\in\phi'(j)}|\bar{\nu}_i(t)-\mu_i|\right)\geq\frac{(\eta-\epsilon)}{2\,c}\right\}\right]\\
&\quad+\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,(L,\phi')\,:I_t=(L,,\phi,\phi'),\,\left(\sum_{j\in L}\sum_{i\in\phi(j)}|\underline{\nu}_i(t)-\mu_i|\right)\geq\frac{(\eta-\epsilon)}{2\,c}\right\}\right]\\
&\leq 8\,N\,\left(\sqrt{2\alpha}+\hat{\Delta}\right)^2\,c^2\,\kappa^2\,m^2\,\sigma^2\,(\eta-\epsilon)^{-2}\,\log T+\frac{8\,N}{T^{\hat{\Delta}^2/2-2}}
\end{aligned}$$

where the final step follows from Corollary 1. Hence combining the bounds for terms (I) and (II) we get the final bound as

$$8\,N\,\kappa^2\,m^2\left(\sqrt{2\alpha}+\hat{\Delta}\right)^2\,c^2\,\sigma^2\,\left(\Delta_{\mathcal{M}^*,q}^{-2}+2\,(\eta-\epsilon)^{-2}\right)\,\log T+\frac{16\,N}{T^{\hat{\Delta}^2/2-2}}+\frac{2\,N\,\alpha}{\alpha-2}$$

and hence we are done. □

Proof of Proposition 5. To address this result we divide the problem based on the decision to accept or reject the null hypothesis which allows us to reduce the problem to an MAB setting in order to apply Corollary 1. We note that

$$\begin{aligned}
R_T&=\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\phi_t\notin\mathbb{F}(\boldsymbol{\mu},\eta)\right\}\right]\\
&=\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\phi_t\notin\mathbb{F}(\boldsymbol{\mu},\eta),\,\delta_t^{\epsilon}=1\right\}\right]+\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\phi_t\notin\mathbb{F}(\boldsymbol{\mu},\eta),\,\delta_t^{\epsilon}=0\right\}\right].
\end{aligned}$$

Now from Lemma C.2, for the second term, we have the bound as

$$8\,N\,\kappa^2\,m^2\left(\sqrt{2\alpha}+\hat{\Delta}\right)^2\,c^2\,\sigma^2\,\left(\Delta_{\mathcal{M}^*,q}^{-2}+(\eta-\epsilon)^{-2}\right)\,\log T+\frac{8\,N}{T^{\hat{\Delta}^2/2-2}}+\frac{2\,N\,(\alpha-1)}{\alpha-2}.$$

For the first term, we note that if $\phi_t\notin\mathbb{F}(\boldsymbol{\mu},\eta)$, which implies that there exists (L,ϕ') such that $g^L(\boldsymbol{\mu};\phi_t\to\phi')<\eta-\Delta_{\mathcal{M}^*,q}$ while $g^L(\underline{\boldsymbol{\nu}}(t);\bar{\boldsymbol{\nu}}(t);\phi_t\to\phi')\geq\epsilon$.
Now as ϵ is chosen carefully, this implies $g^L(\underline{\boldsymbol{\nu}}(t);\bar{\boldsymbol{\nu}}(t);\phi_t\to\phi')\geq g^L(\boldsymbol{\mu};\phi_t\to\phi')$ Hence

$$\begin{aligned}
&\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\phi_t\notin\mathbb{F}(\boldsymbol{\mu},\eta),\,\delta_t^{\epsilon}=1\right\}\right]\\
&\leq\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,(L,\phi')\,:\,g^L(\underline{\boldsymbol{\nu}}(t);\bar{\boldsymbol{\nu}}(t);\phi_t\to\phi')\geq g^L(\boldsymbol{\mu};\phi_t\to\phi')\right\}\right]\\
&\leq\mathbb{E}_{H_a}\left[\sum_{t=1}^{T}\mathbf{1}\left\{\exists\,i\in\mathcal{N}\,:\,\underline{\nu}_i(t)\geq\mu_i\;\text{ or }\mu_i\geq\bar{\nu}_i(t)\right\}\right]\\
&\leq\frac{2\,N\,(\alpha-1)}{\alpha-2}.
\end{aligned}$$

Combining the bounds the result follows. □