

---

# Multi-armed Bandit Algorithm against Strategic Replication

---

Suho Shin\*

Seungjoon Lee<sup>†</sup>

Jungseul Ok<sup>†‡</sup>

\*LINE Plus Corporation

<sup>†</sup>Department of Computer Science and Engineering, Pohang University of Science and Technology

<sup>‡</sup>Graduate School of Artificial Intelligence, Pohang University of Science and Technology

## Abstract

We consider a multi-armed bandit problem in which a set of arms is registered by each agent, and the agent receives reward when its arm is selected. An agent might strategically submit more arms with replications, which can bring more reward by abusing the bandit algorithm’s exploration-exploitation balance. Our analysis reveals that a standard algorithm indeed fails at preventing replication and suffers from linear regret in time  $T$ . We aim to design a bandit algorithm which demotivates replications and also achieves a small cumulative regret. We devise Hierarchical UCB (H-UCB) of replication-proof, which has  $O(\ln T)$ -regret under any equilibrium. We further propose Robust Hierarchical UCB (RH-UCB) which has a sublinear regret even in a realistic scenario with irrational agents replicating careless. We verify our theoretical findings through numerical experiments.

## 1 Introduction

With increasing attention to smart recommender system, multi-armed bandit (MAB) problem is extensively studied under various settings as it captures the fundamental trade-off between exploration and exploitation. In a standard stochastic MAB setting, at each time step, a principal chooses which arm to pull a given set of arms. Once an arm is pulled, a random reward associated to the arm is revealed and the principal receives the reward. The principal’s goal is to maximize the expected cumulative

reward or equivalently to minimize the cumulative regret which is defined by the expected loss compared to oracle policy that always pulls the optimal arm over the given time horizon. To this end, at each decision, the principle needs to address the trade-off between exploring new arms with high potential and exploiting empirically best arms with low risk. A extensive line of works have been studied the fundamental limit in [Lai and Robbins, 1985] and proposed optimal algorithms achieving the limit [Auer et al., 2002, Garivier and Cappé, 2011, Scott, 2010, Chapelle and Li, 2011].

These algorithms work efficiently in ideal case, however, they might suffer from a strategic manipulation of the agents who register the arms in the platform. For example if the principal runs  $\epsilon$ -greedy based algorithm [Auer et al., 2002], the probability of being selected at exploration phase will certainly increase as the agent registers more arms. This implies that the agent should replicate their contents as much as possible to increase their revenue. As a consequence, it will interrupt the platform’s algorithm in identifying the best contents and possibly result in a decrease of the platform’s revenue.

To prevent such abuse of the players, on one hand, the platform can introduce an automated process to detect the duplicated contents and blacklist the owners [AdSense, 2020, Youtube, 2021]. However, these process are often discouraged in some real-world applications since they might be erroneous or too costly to adopt since the platform is required to analyze features of registered contents.

On the other hand, one can borrow some off-the-shelf algorithms proposed in infinitely many-armed bandit literature [Berry et al., 1997, Wang et al., 2008] assuming a worst-case scenario. Yet, there are some limitations on it either: (i) these works usually assume certain characterizations on the pool of arms which might be discouraged in practice, and (ii) the replication of the arms itself could be harmful to the platform since it introduces an additional burden in hardware

resources.

In this context, our major focus is how an algorithm can demotivate the agents in advance from strategically replicating arms, when it can only distinguish to whom each arm belongs.

Our main contributions are summarized as follows:

- (a) To the best of our knowledge, this is the first study to address the issue of strategic replication in multi-armed bandit problem and mathematically analyze it.
- (b) We show that UCB1 is replication-prone and hence admits infinitely many replications for the strategic agents, which obviously consequences linear regret.
- (c) We propose H-UCB which mainly restrict the amount of exploration to be  $O(\ln T)$  for each agent by separating agent and arm selection phase, and prove that it is replication-proof while achieving  $O(\ln T)$ -regret under the equilibrium.
- (d) To cope with practical challenges, we present RH-UCB which integrates arm sampling and enlarges the amount of agent exploration into  $O(\sqrt{T} \ln T)$ , and prove that it has  $O(\sqrt{T} \ln T)$  regret in the presence of replicators who infinitely replicate their arms.
- (e) We provide numerical experiments to support our claim, and present some useful insights from it.

In Section 2, we present our model and objectives. We then formally define replication-proneness with a corresponding negative result of UCB1, and introduce replication-proofness in Section 3. We provide our positive results in 4 and present numerical experiments in Section 5. All the proofs are presented in Appendix C due to space limit.

## 1.1 Related Work

We first present a line of works considering strategic behavior in various multi-armed bandit problem. [Braverman et al., 2019] study a multi-armed bandit problem when each arm itself is a strategic agent, and once the arm is selected by a principal, a random reward is realized, and the agent reports the reward to the principal, and finally this amount of reported reward is delivered to the principal. Here, the reported value can either be genuine or not with respect to the agent’s strategic decision. The agents want to maximize their own cumulative expected utility defined by the summation of undelivered rewards, and the principal aims to extract maximal cumulative reward from the agents. The authors show that existing algorithms maintain a bad approximate equilibrium, and proposed a mechanism that only possesses an efficient approximate equilibrium in which the princi-

pal enjoys sub-linear regret. [Feng et al., 2020] study a similar problem but when the agents try to maximize the total expected number of selections by manipulating their random reward within a fixed amount of total manipulation cost, and show that popular algorithms achieve an intrinsic robustness.

Some of the works consider a strategic behavior of users in user-generated content platforms to motivate the strategic users to promptly contribute their content rather than delaying it [Jain et al., 2014], or to contribute a high quality content which possibly requires their costly effort [Ghosh and Hummel, 2014, Ghosh and McAfee, 2011, Ghosh and Hummel, 2013, Liu and Ho, 2018]. Initiated by [Kremer et al., 2014], there also exists a line of works referred to as incentivized exploration which study incentivizing schemes to motivate myopic selfish agent who selects the arms to explore more arms submitted in the system rather than exploiting the well-seeming ones [Wang and Huang, 2018, Bahar et al., 2020], and also in Bayesian setting [Mansour et al., 2015, Mansour et al., 2016, Frazier et al., 2014]. Since these works are quite distant from our focus of interest, we skip the detailed explanation.

Since we design an algorithm that is robust to the agent’s repetitive replication, it seems worthwhile to present the works studying efficient algorithms in case of infinitely many arms. [Berry et al., 1997] study infinitely many-armed Bernoulli bandit problems when each arm’s parameter is sampled from a common distribution under the assumption that the optimal arm has mean 1. [Wang et al., 2008] and [Carpentier and Valko, 2015] rather assume certain characteristics in near-optimal arms’ reward distribution to attain sublinear algorithms. [Ghalme et al., 2021] consider the scenario where the arms are not fully given at first but become available in a sequential manner, which draws a need of new definition in regret. They show that it is necessary to have certain assumptions on the arrival time of the optimal arm to achieve sublinear regret, and provide corresponding sublinear regret algorithm. Unlike from these approaches, our work mainly focus at preventing such situation by demotivating the agents in advance.

## 2 System model

We consider  $n$  agents, where each agent  $i \in \mathcal{N} = [n] := \{1, 2, \dots, n\}$  is endowed with the set of unique arms, denoted by  $\mathcal{O}_i := \{o_{i,1}, o_{i,2}, \dots, o_{i,l(i)}\}$ , with a constant  $l(i) \geq 1$ . Each arm  $a \in \mathcal{O} := \cup_{i \in \mathcal{N}} \mathcal{O}_i$  is associated with Bernoulli<sup>1</sup> reward distribution of mean  $\mu(a)$ . We con-

<sup>1</sup>This can be extended to the reward distributions of single-parameter exponential family, while it can be corre-

sider a one-shot game scenario with a principal and  $n$  agents, where we assume that agents can hide or replicate a part of its original arms. More formally, in advance of MAB, each agent  $i \in \mathcal{N}$  decides and registers its strategy  $\mathcal{S}_i = \{s_{i,k}^{(c)} : c \in [c_{i,k}], k \in [l(i)]\}$  of support  $\mathcal{O}_i$  with parameters  $(c_{i,1}, c_{i,2}, \dots, c_{i,l(i)})$  such that for  $k \in [l(i)]$ ,  $c_{i,k}$  is the number of copies of  $o_{i,k}$  (including itself) in  $\mathcal{S}_i$  and  $c_{i,k} = 0$  implies that  $o_{i,k}$  is not registered at all, where each copy  $s_{i,k}^{(c)}$  has the same reward distribution of arm  $o_{i,k}$ , i.e.,  $\mu(o_{i,k}) = \mu(s_{i,k}^{(c)})$  for each  $c \in [c_{i,k}]$ .<sup>2</sup> Then, given  $\mathcal{S} := (\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n)$ , at each round  $t = 1, 2, \dots, T$ , the principal selects one of the registered arms, denoted by  $A_t$ , according to an MAB algorithm  $\mathfrak{A}$ , and receives reward  $R_t$  drawn independently from the corresponding distribution of mean  $\mu(A_t)$ . This is indeed a one-shot game as provided the principal's algorithm  $\mathfrak{A}$ , each agent  $i \in \mathcal{N}$  simultaneously decides strategy  $\mathcal{S}_i$  and never updates it once registered. We formally describe the behavior of strategic agents in Section 2.1, and the principal's objective in Section 2.2.

## 2.1 Strategic Agent

For simplicity, we describe strategic agents when agent  $i$  has the utility  $v_i$  defined by the accumulated reward from its arms in strategy  $\mathcal{S}_i$ , i.e.,  $v_i(\mathcal{S}; \mathfrak{A}, T) := \sum_{t=1}^T R_t \mathbb{1}[A_t \in \mathcal{S}_i]$ , which directly corresponds to the canonical case that a fixed portion of the rewards is shared from the principal. We note that the entire analysis can be extended for a more general definition of utility including discount factor over time and non-negative marginal utility, rigorously described in Appendix B. A strategic agent  $i$  aims at maximizing the *expected* utility, denoted by

$$u_i(\mathcal{S}; \mathfrak{A}, T) := \mathbb{E}[v_i(\mathcal{S}; \mathfrak{A}, T)] , \quad (1)$$

where the expectation takes over the randomness of algorithm  $\mathfrak{A}$  and reward  $R_t$ 's. We define *dominant strategy* as follows:

**Definition 1** (Dominant strategy). *Agent  $i$ 's strategy  $\mathcal{S}_i$  is dominant if regardless of the other agents' strategy  $\mathcal{S}_{-i} := (\mathcal{S}_1, \dots, \mathcal{S}_{i-1}, \mathcal{S}_{i+1}, \dots, \mathcal{S}_n)$ , it provides the expected utility at least as much as any other  $\mathcal{S}'_i$  of support  $\mathcal{O}_i$  does, i.e.,*

$$u_i(\mathcal{S}_i, \mathcal{S}_{-i}) \geq u_i(\mathcal{S}'_i, \mathcal{S}_{-i}) , \quad (2)$$

sponded with the plausible scenario of recommender system in which the performance is often measured by the number of total clicks or hits per content.

<sup>2</sup>Such a replication-invariant reward assumption coincides with the time-invariant reward distribution, commonly assumed in literature connecting MAB and recommendation systems, where the time horizon of MAB corresponds to a stream of users with no duplication.

where for simplicity, we often write  $u_i(\mathcal{S}_i, \mathcal{S}_{-i}) = u_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T)$  omitting  $\mathfrak{A}$  and  $T$ . The dominance of  $\mathcal{S}_i$  is strict if the inequality (2) is strict.

It is clear that once an agent  $i$  is able to identify a dominant strategy  $\mathcal{S}_i$ , the agent's rational decision is always selecting  $\mathcal{S}_i$  regardless of the others' strategy  $\mathcal{S}_{-i}$ . Hence, it can be used to characterize the Nash equilibrium of agents' strategy. However, depending on the principal's algorithm and the agent's knowledge, it is possible that there exists no dominant one, or the agent has no ability to identify one. In our analysis, we will elaborate the existence or identifiability of dominant strategy.

## 2.2 Principal's objective

The principal aims at maximizing the utility, which is defined by the accumulated reward from selecting arms submitted by the agents. In other words, the principal's objective can be alternatively formulated as the minimization of the expected cumulative opportunity cost defined as  $\text{Regret}(\mathcal{S}; \mathfrak{A}, T)$  in the followings:

**Definition 2** (Principal's regret). *Given strategy  $\mathcal{S}$ , the regret of algorithm  $\mathfrak{A}$  up to time horizon  $T$  is*

$$\text{Regret}(\mathcal{S}; \mathfrak{A}, T) := \sum_{t=1}^T \sum_{i \in \mathcal{N}} \sum_{a \in \mathcal{S}_i} \Delta(a) \mathbb{E}[\mathbb{1}[A_t = a]] ,$$

where we let  $\Delta(a) := \mu^* - \mu(a)$  denote the expectation of the instantaneous regret from selecting arm  $a$  instead of the optimal arm  $\mu^* := \max_{a' \in \mathcal{O}} \mu(a')$ .

It is well-known that without any prior knowledge, the optimal regret is scaling with the number of arms [Lai and Robbins, 1985]. To be specific, if  $\mathcal{S}$  includes the best arm  $a^* := \arg\max_{a \in \mathcal{O}} \mu(a)$ , then the optimal regret is given by  $O(|\mathcal{S}| \ln T)$ , where  $|\mathcal{S}|$  is the number of arms registered. Such a scaling regret in the number of arms is inevitable mainly because  $O(\ln T)$  explorations per arm is required to information-theoretically identify the best arm  $a^*$  with an error probability less than  $O(1/t)$  at round  $t$  so that the regret from misidentifying  $a^*$  is bounded by  $O(\sum_{t=1}^T 1/t) = O(\ln T)$ . We refer to [Lai and Robbins, 1985] for a rigorous analysis. This illustrates the importance of maintaining the number of suboptimal arms as less as possible, while having smaller number of arms is also beneficial in terms of principal's management cost, e.g., server capacity, in practice.

## 3 Preliminary

In this section, we provide an analysis revealing the vulnerability of a canonical algorithm and a simple

baseline against replicators. This shows not only the significance but also the challenge of the replication problem.

### 3.1 Replication-proneness

We first investigate the strategic behavior of agents when using a canonical algorithm of regret minimization, and check if the agents have a strong motivation to replicate arms under the algorithm, i.e., the algorithm is *replication-prone*.

**Definition 3** (Replication-proneness). *An algorithm  $\mathfrak{A}$  for time horizon  $T$  is replication-prone if for an agent  $i \in \mathcal{N}$ , any strategy  $\mathcal{S} = (\mathcal{S}_i, \mathcal{S}_{-i})$ , there exists  $\mathcal{S}'_i$  such that  $\mathcal{S}_i \subset \mathcal{S}'_i$  and*

$$u_i(\mathcal{S}'_i, \mathcal{S}_{-i}; \mathfrak{A}, T) > u_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T). \quad (3)$$

In words, a replication-prone algorithm provides a strictly positive gain from replicating a part of arms for each agent, and thus each agent's best decision is to register infinitely many arms regardless of the others' strategy. Our analysis verifies that UCB1 [Auer et al., 2002] is replication-prone, meanwhile we believe that an extended analysis can show the replication-proneness of other canonical MAB algorithms such as KL-UCB [Garivier and Cappé, 2011].

**Theorem 1.** *UCB1 is replication-prone.*

This theorem shows a theoretical understanding of replicating behaviors in practice, and an importance to prevent replicators. It is intuitive from the fact that UCB1 is oblivious to agents and thus grants a certain amount of explorations for every single registered arm. We provide a formal proof of Theorem 1 in Appendix C, where we mainly use a sophisticated coupling argument to compare rewards from two MAB processes with or without an additional replication of the entire set of agent  $i$ 's original arms. We note that our analysis requires no special condition on  $\mathcal{O}$  to show (3), and thus just given UCB1, every agent  $i$  is able to realize that its best decision must include infinitely many replications even without knowledge on its arms or others.

As mentioned earlier, UCB1 with registration of sub-optimal arms more than  $\Omega(T/\ln T)$  may suffer from linear regret. To overcome such an unwilling behavior of the algorithm, one might consider using the algorithms provided in many-armed bandit literature. However, as we briefly discussed in Section 1, this may not work well when only some of the agents replicate their arm, and even the replication itself harms the principal since it exhausts the hardware resources. Therefore, we instead aim at demotivating the agents replicating the arms in advance.

### 3.2 Replication-proofness and Fair( $\mathfrak{A}$ )

To prevent the infinite replicas in advance, we focus on constructing a mechanism that makes agents' Nash equilibrium at not replicating the arms. To this end, we define *replication-proof* algorithm as a solution concept of our problem as follows:

**Definition 4** (Replication-proofness). *An algorithm  $\mathfrak{A}$  is replication-proof if there exists a dominant strategy that no agent replicates.*

Due to the equality in (2) the replication-proofness actually refers to the weakly dominant strategy in the literature. Hence, it does not guarantee that making no replication is making more money for the agents, which possibly induces a deviation for the agents. However, in practice, this can be prevented by introducing a tiny constant cost in registering the arms or charging a periodical commission fee with respect to the number of registered arms in the platform. In this sense, under a replication-proof algorithm, the agents have a Nash equilibrium without replicating arms. Replication-proofness and -proneness are not complementary to each other, while one implies the negation of the other, i.e., UCB1 is replication-prone and thus it is not replication-proof. We remark that Definition 4 shares a similar virtue with the truthfulness in the mechanism design literature [Roughgarden, 2010] since truthfully reporting a private valuation can be regarded as reporting only a part of the original arms without any replication in our scenario.

We note that the replication-proofness can be achieved by a very simple baseline equipped with any MAB algorithm  $\mathfrak{A}$ , denoted by Fair( $\mathfrak{A}$ ). At each round, Fair( $\mathfrak{A}$ ) chooses an agent  $i \in \mathcal{N}$  uniformly at random, and then plays an arm  $a \in \mathcal{S}_i$  from running  $\mathfrak{A}$  based on the history of the agent  $i$ 's arms. Then it is straightforward to check that Fair( $\mathfrak{A}$ ) is replication-proof since the selection of each agent is not affected by the other agents' strategy for any  $\mathfrak{A}$ . In addition, if possible, the best strategy of each agent is to register its best arm. However, even with the best strategy containing only each agent's best arm, every agent is selected with the equal probability at each round, and thus any suboptimal agents will be selected in linear portion of rounds, i.e., linear regret is inevitable.

Therefore, the principal's objective can be given in two folds: i) demotivating the agents from replicating arms; and ii) acquiring small regret. More formally, we aim at designing an algorithm that is replication-proof and has sublinear regret at the same time.

## 4 Main result

In what follows, we first describe an replication-proof algorithm, named Hierarchical UCB (H-UCB), which has a sublinear regret assuming every agent is fully rational and thus not replicating arms. We then propose a robust version of H-UCB (RH-UCB), that are guaranteed to be replication-proof and of sublinear regret even in the realistic scenario possibly with irrational agents.

### 4.1 Hierarchical UCB

We recall that Fair( $\mathcal{A}$ ) is equipped with the replication-proofness by the agent selection, irrelevant to the number of arms that the agent has, in advance of the arm selection. Inspired by such a hierarchy of selections of agent and then arm, we devise hierarchical UCB (H-UCB) consisting of two phases per round, where **Phase 1** to choose an agent is followed by **Phase 2** to select its arm, sequentially. We provide a pseudo code of H-UCB in Algorithm 1, where we keep track of the empirical mean rewards from each agent  $i$  and arm  $a$  by  $R(i)$  and  $r(a)$ , respectively, and the numbers of rounds selecting agent  $i$ 's arms and an arm  $a$  by  $N(i)$  and  $n(a)$ , respectively. In **Phase 1**, as a part of identifying an agent with the optimal arm, a UCB algorithm selects an agent  $\hat{i}$  of the highest value of agent index  $R(i) + \sqrt{\frac{2 \ln t}{N(i)}}$  which estimates the potential rewards from agent  $i$ 's arms at  $t$ -th round. Then, in **Phase 2**, given the empirically best agent  $\hat{i}$ , we aim at selecting the agent  $\hat{i}$ 's best arm. To this end, another UCB algorithm is employed to choose an arm  $\hat{a}$  from the agent  $\hat{i}$  of the highest value of arm index  $r(a) + \sqrt{\frac{2 \ln N(\hat{i})}{n(a)}}$  that assesses the potential of selecting arm  $a$  on the part of time horizon when selecting the agent  $\hat{i}$  only. This forms a hierarchy of UCB algorithms as the arm selection in **Phase 2** depends on the agent selection in **Phase 1**. We then have the following analysis on H-UCB:

**Theorem 2.** *H-UCB is replication-proof.*

Intuitively, the exploration of the best agent in *phase 1* distributes  $O(\ln t)$  budget for each agent regardless of the number of the agent's arms. In addition, given the asymptotically equal budget per agent, if the agent replicated its arms, then it would increase the risk of misidentifying its best arm and reduce its utility. Hence, this demotivates the agent from replicating arms. A formal proof of Theorem 2 is provided in Appendix C. In the proof, we in fact show a stronger notion of replication-proofness based on stochastic orderings to compare two different strategies instance-wisely. Thus, H-UCB is replication-proof even when agent seeks gain from replication in randomness.

---

### Algorithm 1: Hierarchical UCB (H-UCB)

---

```

1 Initialize  $R(i), N(i), r(a)$  and  $n(a)$  at zero for all
    $i \in \mathcal{N}$  and  $a \in \cup_{i \in \mathcal{N}} \mathcal{S}_i$ ;
2 for  $t = 1, 2, \dots$  do
   // Phase 1 - agent selection
3   if Unexplored agent exists then
4     | Pick an unexplored  $\hat{i} \in \{i \in \mathcal{N} : N(i) = 0\}$ ;
5   else
6     | Pick  $\hat{i} = \operatorname{argmax}_{i \in \mathcal{N}} \left\{ R(i) + \sqrt{\frac{2 \ln t}{N(i)}} \right\}$ ;
7   end
   // Phase 2 - arm selection
8   if Unexplored arm exists in  $\mathcal{S}_{\hat{i}}$  then
9     | Pick an unexplored
    $\hat{a} \in \{a \in \mathcal{S}_{\hat{i}} : n(a) = 0\}$ ;
10  else
11    | Pick  $\hat{a} = \operatorname{argmax}_{a \in \mathcal{S}_{\hat{i}}} \left\{ r(a) + \sqrt{\frac{2 \ln N(\hat{i})}{n(a)}} \right\}$ ;
12  end
13  Play  $A_t = \hat{a}$  and receive reward  $R_t$ ;
14  Update statistics:
    $r(\hat{a}) \leftarrow \frac{r(\hat{a})n(\hat{a}) + R_t}{n(\hat{a}) + 1}$ ;  $R(\hat{i}) \leftarrow \frac{R(\hat{i})N(\hat{i}) + R_t}{N(\hat{i}) + 1}$ ;
15   $n(\hat{a}) \leftarrow n(\hat{a}) + 1$ ;  $N(\hat{i}) \leftarrow N(\hat{i}) + 1$ ;
16 end

```

---

We remark that under H-UCB, assuming each agent  $i$  knows its best arm  $a_i^* \in \operatorname{argmax}_{a \in \mathcal{O}_i} \mu(a)$ , the agent's dominant strategy is to register only  $a_i^*$ 's. On such an equilibrium, we obtain a logarithmic bound on the regret of H-UCB:

**Theorem 3.** *Given the original arms  $\mathcal{O}$  and the corresponding mean rewards  $\{\mu(a) : a \in \mathcal{O}\}$ , under an equilibrium  $\mathcal{S}$  when each agent  $i$  decides its strategy knowing its best arm  $a_i^*$ ,  $\operatorname{Regret}(\mathcal{S}; \text{H-UCB}, T)$  is at most*

$$\sum_{i: \Delta_i > 0} \frac{8}{\Delta_i} \ln T + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{i=1}^n \Delta_i\right),$$

where  $\Delta_i := \mu^* - \max_{a \in \mathcal{O}_i} \mu(a)$ .

The proof of Theorem 3 is given in Appendix C, where we present a formal characterization of equilibrium under the assumption. We note that the asymptotic order of regret bound is  $O(n \ln T)$ , which scales with the number of agents rather than the total number of arms, while in practice, it is much harder to replicate agents than arms since a genuine identification, such as social security number or bank account, is required to register an agent.

## 4.2 Robust Hierarchical UCB

The logarithmic regret of H-UCB in Theorem 3 assumes that every agent is rational enough to understand no benefit from replicating arms under H-UCB, and registers its arm truthfully. Meanwhile, in practice, it is still possible to have *replicators* who carelessly duplicate their arms even under a replication-proof algorithm. We assume that each replicator  $i$ 's strategy  $\mathcal{S}_i$  includes at least a constant portion of arms of its best mean reward  $\mu_i^* := \max_{a \in \mathcal{O}_i} \mu(a)$ , i.e., there exists a constant  $c > 0$  such that  $\min_{i \in \mathcal{N}} \frac{|\{a \in \mathcal{S}_i: \mu_i^* = \mu(a)\}|}{|\mathcal{S}_i|} > c$ . This is a reasonable assumption since no replicator wants to hide its best arm. Even when a replicator cannot identify its best arm, the safest choice is to replicate the entire arms equally.

To obtain sublinear regret even with the presence of such replicators, modifying H-UCB, we propose robust H-UCB (RH-UCB), described in Algorithm 2. Before explaining the design of RH-UCB, we note that RH-UCB inherits the hierarchical structure of H-UCB. Hence, an analog analysis to that of Theorem 2 concludes that:

**Theorem 4.** *RH-UCB is replication-proof.*

We now explain rationales behind RH-UCB, where we append a modification to each phase of H-UCB. In **Phase 2**, for each agent  $i$ , the risk of misidentifying the best arm in strategy  $\mathcal{S}_i$  increases with the size of  $\mathcal{S}_i$ . Hence, as a part of bounding the risk in **Phase 2**, RH-UCB begins with  $\mathcal{B}_i$ 's from subsampling each  $\mathcal{S}_i$  uniformly at random without replacement such that the cardinality of subset  $\mathcal{B}_i$  is at most  $L \ln T$ , where  $L > 0$  is a hyperparameter. The choice of size bound depends on  $T$  for analytical simplicity, while our analysis can be extended to an online subsampling with increasing size bound  $L \ln t$ . Given the subsampled  $\mathcal{B}_i$ 's, in **Phase 2**, we use the same UCB1 for  $\mathcal{B}_i$ . Although the number of arms in  $\mathcal{B}_i$  has the upper bound  $L \ln T$ , but it is scaling in  $T$ . Hence, the exploration budget  $O(\ln T)$  distributed over agents in H-UCB may be insufficient to identify each agent's best arm at a desired confidence. For this reason, we increase confidence interval of **Phase 1** from  $\sqrt{\frac{\ln t}{N(i)}}$  to  $\sqrt{\frac{\sqrt{t} \ln t}{N(i)}}$ .

For ease of exposition, we assume that there exists a unique agent  $i^*$  who possesses the optimal arm<sup>3</sup>, and present the following regret bound for RH-UCB:

**Theorem 5.** *Given the original arms  $\mathcal{O}$  and  $\{\mu(a) : a \in \mathcal{O}\}$ , under any  $\mathcal{S}$  given that all the agents are either strategic or replicator, if  $L \geq 1/c$ , then*

<sup>3</sup>Indeed, multiple optimal agents only reduce the regret bound, since the constant  $C$  in our regret bound  $O(C\sqrt{T} \ln T)$  will decrease and  $O(\ln^2 T)$  appear instead.

Regret( $\mathcal{S}; \text{RH-UCB}, T$ ) is at most

$$\sum_{i \in \mathcal{N}} \sqrt{T} \ln T \left( 1 + \frac{4}{\Delta_i^2} + \sum_{\substack{a \in \mathcal{O}_{i^*} \\ \delta_{i^*, a} > 0}} \frac{50L}{\delta_{i^*, a}} + \sum_{\substack{a \in \mathcal{O}_i \\ \delta_{i, a} > 0}} \frac{114L}{\delta_{i, a}} \right) + O(\ln^5 T),$$

where  $\delta_{i, a} = \mu_i^* - \mu(a)$ .

---

### Algorithm 2: Robust H-UCB (RH-UCB)

---

```

1 Initialize  $R(i), N(i), r(a)$  and  $n(a)$  at zero for all
    $i \in \mathcal{N}$  and  $a \in \cup_{i \in \mathcal{N}} \mathcal{S}_i$ ;
2 Subsample  $\mathcal{B}_i$  of size  $\min\{|\mathcal{S}_i|, L \ln T\}$  from  $\mathcal{S}_i$ ;
3 for  $t = 1, 2, \dots$  do
   // Phase 1 - agent selection
4   if Unexplored agent exists then
5     | Pick an unexplored  $\hat{i} \in \{i \in \mathcal{N} : N(i) = 0\}$ ;
6   else
7     | Pick  $\hat{i} = \operatorname{argmax}_{i \in \mathcal{N}} \left\{ R(i) + \sqrt{\frac{\sqrt{t} \ln t}{N(i)}} \right\}$ ;
8   end
   // Phase 2 - arm selection
9   if Unexplored arm exists in  $\mathcal{B}_{\hat{i}}$  then
10    | Pick an unexplored
        $\hat{a} \in \{a \in \mathcal{B}_{\hat{i}} : n(a) = 0\}$ ;
11  else
12    | Pick  $\hat{a} = \operatorname{argmax}_{a \in \mathcal{B}_{\hat{i}}} \left\{ r(a) + \sqrt{\frac{2 \ln N(\hat{i})}{n(a)}} \right\}$ ;
13  end
14  Play  $A_t = \hat{a}$  and receive reward  $R_t$ ;
15  Update statistics:
        $r(\hat{a}) \leftarrow \frac{r(\hat{a})n(\hat{a}) + R_t}{n(\hat{a}) + 1}$ ;  $R(\hat{i}) \leftarrow \frac{R(\hat{i})N(\hat{i}) + R_t}{N(\hat{i}) + 1}$ ;
        $n(\hat{a}) \leftarrow n(\hat{a}) + 1$ ;  $N(\hat{i}) \leftarrow N(\hat{i}) + 1$ ;
16 end
    
```

---

In Appendix C, we provide not only the proof of Theorem 5 but also a further discussion on each term in the regret bound. Even with the replicators, the regret of RH-UCB is asymptotically  $O(\sqrt{T} \ln T)$  for any given configuration of the original arms. Hence, RH-UCB is replication-proof but also achieves sublinear. The majority of regret  $O(\sqrt{T} \ln T)$  is from the probability of misidentifying each agent  $i$ 's best arm among subsampled  $\mathcal{B}_i$  of size  $O(\ln T)$ . We note that  $O(\sqrt{T} \ln T)$  amount of agent exploration indeed optimizes the order of regret bound in our analysis. For any 1-dimensional increasing function  $f$ , suppose that **Phase 1**'s agent exploration term is given by  $R_i + \sqrt{\frac{f(t)}{N(i)}}$ . Then, it can similarly be derived that regret bound becomes  $O(f(T)) + O(\frac{T \ln^2 T}{f(T)})$ , where the optimized regret can be obtained for  $f(t) = O(\sqrt{t} \ln t)$  due to the inequality of arithmetic and geometric means.

We note that the regret analysis in Theorem 5 is problem-dependent as the bound is given as a function

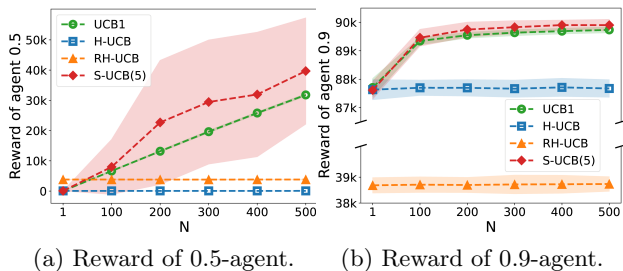


Figure 1: (a) The average revenue and one-sigma interval for 0.5-agent as it replicates more arms in the single original arm setup. (b) The revenue of 0.9-agent in the same setup with (a) but when 0.9-agent replicates.

of the configuration of original arms. In the following theorem, we further present a problem-independent regret bound of RH-UCB:

**Theorem 6.** *RH-UCB has problem-independent regret upper-bound  $O(\frac{n}{\sqrt{c}}T^{\frac{3}{4}}\ln T)$ .*

The detailed terms and proof are presented in Appendix C, where we provide a worst-case analysis on regret. We note that with no replication, the optimal problem-independent regret is  $O(\sqrt{T})$  [Audibert et al., 2009], while RH-UCB has  $O(T^{3/4}\ln T)$  even with replicators. The gap between  $O(\sqrt{T})$  and  $O(T^{3/4}\ln T)$  can be interpreted as an algorithmic cost for being robust and replication-proof.

Note that our regret bound holds for  $L \geq 1/c$ . Running RH-UCB with sufficiently large  $L$  would guarantee this bound, but it might hurt the algorithm’s efficiency since there could be a chance that too much redundant arms are subsampled for each agent. In case when precisely estimating  $c$  a priori is difficult, we can instead subsample  $\ln^2 T$  arms per agent since sufficiently large  $T$  would guarantee  $\ln T \geq 1/c$ . Meanwhile, this still requires the principal to determine  $T$  in advance. This dependency can be removed by gradually subsampling arms up to  $\ln^2 T$  within each agent. These techniques indeed enable us to derive  $O(\frac{n}{\sqrt{c}}\sqrt{T}\ln^3 T)$  regret and  $O(\frac{n}{\sqrt{c}}T^{\frac{3}{4}}\ln^{\frac{3}{2}} T)$  problem-independent regret without any prior information on  $L$  and  $T$ , where we put back a detailed description of the algorithm and corresponding analysis in Appendix D.

We remark that both the hierarchical structure and the subsampling are necessary to ensure sublinear regret. If there exists no hierarchy, it might work poor when only the suboptimal agents are replicators, but optimal one is not. Otherwise if subsampling is discarded, any replicators would hurt the system since their optimal arms cannot be identified. We provide a relevant ablation study in Section 5.

Finally, it is obvious that H-UCB works better when all the agents only register their optimal arm, since RH-UCB achieves  $O(\sqrt{T}\ln T)$ -regret in this case. However, we need to put more exploration in Phase 1 to ensure that all the arms within the agent are explored sufficiently under the existence of replicators. Hence, by closely observing behavior of the agents in the system, one may balance the trade-off between robustness and efficiency. We will present a relevant case study in the following section.

## 5 Simulation

We evaluate the proposed algorithms and compare them with existing ones by borrowing well-known open-source library SMPyBandits [Besson, 2018]. In addition to UCB1, H-UCB and RH-UCB, we measure the performance of S-UCB( $l$ ) which samples at most  $l \ln T$  arms among the entire arms given  $T$ , and then runs UCB1 on them. That is, S-UCB( $l$ ) is an algorithm where the hierarchy is removed from RH-UCB, discarding the agent selection phase. We provide a pseudo-code in Appendix E. We decide the subsampling ratio  $l$  as the number of total original arms. Each solid or dashed line indicates the mean value of the treatment over the repetitions, and its shade refers to the one-sigma interval.

Figure 1(a) and 1(b) represent the changes of agents’ revenue with respect to their strategies in each algorithm. In both scenarios, we consider 5 agents each of whom has a single original Bernoulli arm with parameters 0.5, 0.6, 0.7, 0.8, 0.9, respectively. We denote  $x$ -agent as the agent whose optimal arm’s parameter is  $x$ . Figure 1(a) refers to 0.5-agent’s revenue when only the 0.5-agent replicates its arm among the agents, and Figure 1(b) indicates that of 0.9-agent when only the 0.9-agent replicates.

### Vulnerability of non-hierarchical algorithms.

The replicating agents’ mean revenue increases in both scenarios under UCB1 and S-UCB, but stays consistent under H-UCB and RH-UCB. The revenue increment of 0.5-agent is more drastic than that of 0.9-agent. It is because the replication of the optimal arm does not fundamentally change the asymptotic selection ratio of any sub-optimal arms, but only delays their selections. Note that the 0.5-agent’s revenue increases almost in linear manner which is quite intuitive regarding that every suboptimal arm is sampled  $O(\ln T)$  times in UCB1. On the other hand, in S-UCB, the revenue increases in somewhat convex manner due to the nature of probabilistic subsampling.

**Hierarchy makes low regret.** Figure 2(a) and Figure 2(b) depicts the principal’s cumulative regret in

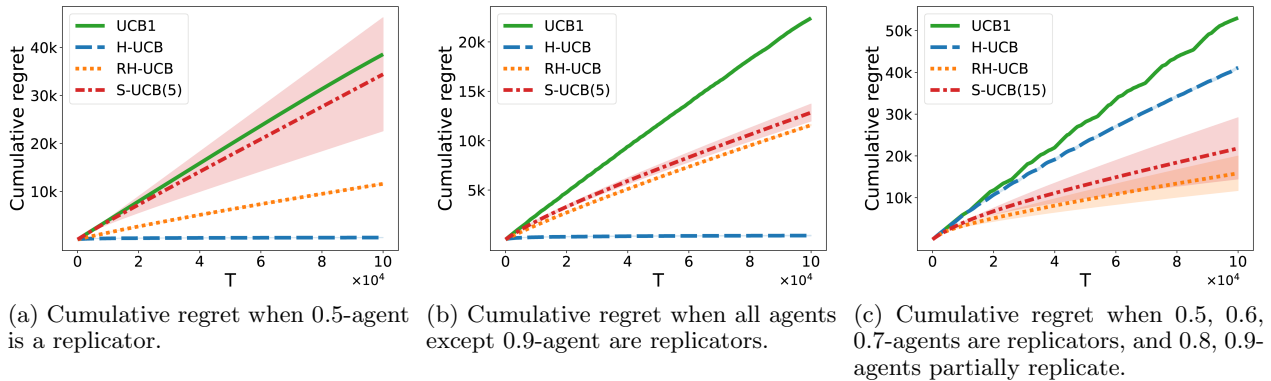


Figure 2: (a) The average cumulative regret and one-sigma interval for the principal in each algorithm, when 0.5-agent replicates its arm 1000 times in the single original arm setup. (b) This shares the same setup with (a), but when all the agents except 0.9-agent replicate their arms 1000 times. (c) 0.5, 0.6, 0.7-agents replicate their original arms 1000 times and 0.8, 0.9-agents partially replicate their arms in the multiple original arms setup.

the single original Bernoulli arm setup with 5 agents like Figure 1. Figure 2(a) shows the regret when the 0.5-agent replicates its arm 1000 times but the others do not replicate at all. UCB1 and S-UCB suffer from the replication even if there is only one replicator agent. However, the hierarchy-based ones allocate a fixed portion of exploration within each agent. Because that amount of exploration is shared across the agent’s arms, the replication does not significantly harm the principal. Since RH-UCB enforces a larger amount of agent exploration, low-reward agents are also explored more than H-UCB. This results in worse performance for RH-UCB than H-UCB.

Besides, in Figure 2(b), we consider a similar scenario but where 0.5, 0.6, 0.7, 0.8-agents replicate their arms 1000 times yet only the 0.9-agent does not. H-UCB still achieves lower regret than RH-UCB since the optimal agent does not replicate arms at all. Hence we believe that H-UCB might work well if there exists any optimal agent who does not replicate its arm at all. This indeed can be observed in our analysis, where the predominant term  $O(L\sqrt{T}\ln T)$  can be removed in the regret bound in Theorem 5.

### Primacy of RH-UCB in a world of replicators.

On the other hand, Figure 2(c) represents the regime that RH-UCB works more robustly. We assume that there are 5 agents with 3 Bernoulli original arms. Each of those agents has arms with mean 0.2 and 0.1 in common, and has a different optimal arm whose parameter is 0.5, 0.6, 0.7, 0.8, or 0.9, respectively. The 0.5, 0.6, and 0.7-agents are replicators and they register 1000 arms per each original arm. Besides, we assume that the 0.8-agent and the 0.9-agent are *partial replicators*. Each of those two agents registers its own optimal arm 10 times, and the suboptimal arms 100 times each. In this case, the power of RH-UCB reveals so that it suc-

ceeds to identify the best agent and the corresponding best arm. However, the regret of H-UCB grows almost in linear manner, which is even worse than S-UCB. This is because H-UCB only explores each agent  $O(\ln T)$  times, and the existence of many arms in the optimal agent possibly hinders H-UCB to discover the best arm within the agent. This would be the case when all the agents are replicators.

Our intention of designing hierarchically-structured algorithms is solely to demotivate the agents from replicating their arms. Interestingly, however, our algorithms outperform existing ones in the context of cumulative regret in many regimes. This might arise because there is an inherent dominance structure among the agents in the setups of our experiments including that of Figure 2(c). In the setups, the suboptimal arms remain the same across the agents, and the agents are only distinguished by the optimal arm of each agent so that the parameter of the optimal arm fully determines an agent-wise dominance. Our hierarchical algorithms naturally exploit such inherent dominance structure among the registered arms by grouping them by agent, which possibly increases their performance. Hence, we believe that H-UCB or RH-UCB would perform better than our theoretical guarantees since many practical scenarios may possess the structure across the agents.

## 6 Conclusion

We analytically study multi-armed bandit algorithms against strategic replication. We firstly show that UCB1 admits an infinitely many registration of arms for strategic agents which inevitably induces linear regret. We mainly observe that it is essential to limit the amount of exploration per agent to demotivate such strategic replication, and present that a simple fair al-



gorithm achieves it while it suffers from linear regret. We propose H-UCB to catch the both rabbits of sublinear regret and replication-proofness, and present more robust algorithm named RH-UCB which adapts to the infinitely many replicas from replicators. We conduct numerical experiments to validate our theoretical findings, and provide some practical insights from it.

### Acknowledgements

This work was supported by NRF (No. 2021M3E5D2A01023887) and IITP (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) grant funded by the Korea government (MSIT).

### References

- [AdSense, 2020] AdSense, G. (2020). Youtube adsense replication policies. [https://www.youtube.com/watch?v=F-sX\\_XRdtTM](https://www.youtube.com/watch?v=F-sX_XRdtTM).
- [Amin et al., 2013] Amin, K., Rostamizadeh, A., and Syed, U. (2013). Learning prices for repeated auctions with strategic buyers. *arXiv preprint arXiv:1311.6838*.
- [Audibert et al., 2009] Audibert, J.-Y., Bubeck, S., et al. (2009). Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122.
- [Auer, 2002] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.
- [Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.
- [Bahar et al., 2020] Bahar, G., Ben-Porat, O., Leyton-Brown, K., and Tennenholtz, M. (2020). Fiduciary bandits. In *International Conference on Machine Learning*, pages 518–527. PMLR.
- [Berry et al., 1997] Berry, D. A., Chen, R. W., Zame, A., Heath, D. C., and Shepp, L. A. (1997). Bandit problems with infinitely many arms. *The Annals of Statistics*, pages 2103–2116.
- [Besson, 2018] Besson, L. (2018). SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python. Online at: [github.com/SMPyBandits/SMPyBandits](https://github.com/SMPyBandits/SMPyBandits). Code at <https://github.com/SMPyBandits/SMPyBandits/>, documentation at <https://smpybandits.github.io/>.
- [Braverman et al., 2019] Braverman, M., Mao, J., Schneider, J., and Weinberg, S. M. (2019). Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pages 383–416. PMLR.
- [Carpentier and Valko, 2015] Carpentier, A. and Valko, M. (2015). Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141. PMLR.
- [Chapelle and Li, 2011] Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257.
- [Combes et al., 2017] Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. *arXiv preprint arXiv:1711.00400*.
- [Feng et al., 2020] Feng, Z., Parkes, D., and Xu, H. (2020). The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning*, pages 3092–3101. PMLR.
- [Frazier et al., 2014] Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. (2014). Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22.
- [Garivier and Cappé, 2011] Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.
- [Ghalme et al., 2021] Ghalme, G., Dhamal, S., Jain, S., Gujar, S., and Narahari, Y. (2021). Ballooning multi-armed bandits. *Artificial Intelligence*, 296:103485.
- [Ghosh and Hummel, 2013] Ghosh, A. and Hummel, P. (2013). Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 233–246.
- [Ghosh and Hummel, 2014] Ghosh, A. and Hummel, P. (2014). A game-theoretic analysis of rank-order mechanisms for user-generated content. *Journal of Economic Theory*, 154:349–374.
- [Ghosh and McAfee, 2011] Ghosh, A. and McAfee, P. (2011). Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, pages 137–146.

- [Jain et al., 2014] Jain, S., Chen, Y., and Parkes, D. C. (2014). Designing incentives for online question-and-answer forums. *Games and Economic Behavior*, 86:458–474.
- [Jun et al., 2019] Jun, K.-S., Willett, R., Wright, S., and Nowak, R. (2019). Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pages 3163–3172. PMLR.
- [Katariya et al., 2017] Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017). Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, pages 392–401. PMLR.
- [Kremer et al., 2014] Kremer, I., Mansour, Y., and Perry, M. (2014). Implementing the “wisdom of the crowd”. *Journal of Political Economy*, 122(5):988–1012.
- [Kveton et al., 2017] Kveton, B., Szepesvári, C., Rao, A., Wen, Z., Abbasi-Yadkori, Y., and Muthukrishnan, S. (2017). Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*.
- [Lai and Robbins, 1985] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.
- [Liu and Ho, 2018] Liu, Y. and Ho, C.-J. (2018). Incentivizing high quality user contributions: New arm generation in bandit learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [Mansour et al., 2015] Mansour, Y., Slivkins, A., and Syrgkanis, V. (2015). Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582.
- [Mansour et al., 2016] Mansour, Y., Slivkins, A., Syrgkanis, V., and Wu, Z. S. (2016). Bayesian exploration: Incentivizing exploration in bayesian games. *arXiv preprint arXiv:1602.07570*.
- [Mohri and Medina, 2014] Mohri, M. and Medina, A. M. (2014). Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *International Conference on Machine Learning*, pages 262–270. PMLR.
- [Roch, 2015] Roch, S. (2015). Modern discrete probability: An essential toolkit. *University Lecture*.
- [Roughgarden, 2010] Roughgarden, T. (2010). Algorithmic game theory. *Communications of the ACM*, 53(7):78–86.
- [Scott, 2010] Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.
- [Wang and Huang, 2018] Wang, S. and Huang, L. (2018). Multi-armed bandits with compensation. *arXiv preprint arXiv:1811.01715*.
- [Wang et al., 2008] Wang, Y., Audibert, J.-Y., and Munos, R. (2008). Infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*.
- [Youtube, 2021] Youtube, G. (2021). Youtube channel monetization policies. [https://support.google.com/youtube/answer/1311392?hl=en&ref\\_topic=9153642](https://support.google.com/youtube/answer/1311392?hl=en&ref_topic=9153642).

---

## Supplementary Material: Multi-armed Bandit Algorithm against Strategic Replication

---

### A Open Problems and Future Directions

We discuss some open questions and future research directions in this section.

**Tightening regret bound.** We show that there exists an algorithm that is replication-proof and also has a sublinear regret regardless of whether agents are strategic or not. However, it is not studied whether the regret bound we obtain is tight or not. To tighten the regret analysis, a fundamental limit analysis is necessary. To be specific, we need a regret lower bound for uniformly good and also replication-proof algorithms. The fundamental analysis with the notion of replication-proofness seems challenging. Hence, we would bypass it by studying a lower bound for uniformly good algorithms given the scenario of replicator in our paper with an argument of change-of-measure, e.g., [Combes et al., 2017]. The relaxed scope of algorithms of interest would provide a loose lower bound to be achieved by replication-proof algorithms. Meanwhile, this can provide a useful insight to improve the regret upper bound fully exploiting the assumption on replicators, i.e.,  $\min_{i \in \mathcal{N}} \frac{|\{a \in \mathcal{S}_i : \mu_i^*(a) = \mu(a)\}|}{|\mathcal{S}_i|} > c > 0$ , as in [Combes et al., 2017], the lower bound analysis corresponds to an optimal upper bound algorithm. In fact, the assumption is analog to the bandit problems with linear or row-rank structures [Auer, 2002, Jun et al., 2019, Katariya et al., 2017, Kveton et al., 2017]. However, our bandit problem with replicators is more challenging since we do not know an embedding (or feature map) of arms, which are provided in the linear and low-rank bandit problems.

**Agent information model.** We assume that the strategic agents have ability to *exactly* assess their arms, or at least to identify the best one. This may be not true in practice. Hence, one can consider a Bayesian information model where each agent only knows (or believes) the meta distribution of its arm’s parameter, but does not know the arm’s distribution exactly. This can be interpreted as a scenario of recommender system to a population of heterogeneous clients. Perhaps this meta distribution could be different for each arms within the agent. In these cases, the agents might be motivated to register all the original arms since it will suffer from linear regret with constant probability if some of the original arms are not registered. This is somewhat different from our result, since our model encourages the agent to exactly register only the best arm if such information is readily available. This difference seems natural since any agent who is oblivious to its own arms would definitely register all the arms if there exists no significant cost in registering more arms. Nonetheless, in both cases, we believe that our hierarchical algorithms would also demotivate the agents from replicating their arms since replication might not result in an increase of the expected value of the empirical average reward.

**Beyond UCB-based algorithms.** Although our analysis on the negative result, i.e. replication-proneness, is shown for the case of UCB1, we believe that our proof can be extended to other types of canonical algorithms, e.g.  $\varepsilon$ -greedy based algorithms [Auer et al., 2002] or Thompson Sampling [Scott, 2010, Chapelle and Li, 2011]. One can alternatively run some simple numerical experiments to verify that the agents’ reward indeed increase by replicating the arms as well in these types of algorithms. Regarding our positive results, we mainly reveal that an algorithm needs to restrict its exploration cost for each agent to achieve replication-proofness. For  $\varepsilon$ -greedy based algorithms, it is straightforward to examine the exploration cost since the exploration phase and exploitation phase are explicitly separated. Hence, we believe that the hierarchical structure can analogously be applied to construct a replication-proof and sublinear regret  $\varepsilon$ -greedy based algorithm. Meanwhile, it becomes nontrivial for algorithms based on probability matching including Thompson Sampling. The exploration and exploitation phase are not clearly separated in these algorithms, hence we need to define the exploration cost in rigorous manner so that restricting exploration cost would result in a replication-proofness. These analysis might be more challenging since we need to perform Bayesian analysis, and even these algorithms might be discouraged since they need information on a parametric structure of reward distributions. However, we believe that there may exist some valuable insights in these analysis, since different notion of exploration cost could lead to a different point of the equilibrium. This might bridge the gap between our notion of replication-proofness

and replication-proneness. Indeed, if we construct a logarithmic regret algorithm that is not replication-prone, but only admits a finite number of replications to the strategic agents, then it could be better than RH-UCB for the principal since the finite number of replications only contributes to increase constant term in the regret bound while RH-UCB admits square root regret. To fairly compare the various algorithms in this context, it might require us to define a new criterion to estimate their efficiency, e.g., worst-case regret over the dominant strategies (or over any strategies if no such one exists).

**Beyond stochastic bandit with one-shot game.** In practice, many recommendation systems are modeled as contextual bandit which utilizes features of clients and arms. Since replicating could still hurt the platform under the contextual bandit algorithm, we believe that we can generalize our results into this setting. Besides, we can consider a sequential arrival of the agents, which will result in a repeated-game between the principal and the agents. In this case, the platform might have more controllability on handling agent’s strategic behavior since agents are forced to register only certain amount of contents at once, and principal can block the agent’s registration in advance at some rounds if they’ve been registered too much till that time. On the other hand, agents could also abuse the system more viciously, for example, they can get rid of their registered contents, and register it again as if it is newly created ones, which might hamper the algorithm from identifying the optimal arm. Finally we remark that our analysis on replication-proofness are mainly based on the fact that a family of Bernoulli distributions equipped with stochastic ordering form a totally ordered set. If such ordering cannot be made within a family of distributions in consideration, then we cannot compare utilities between two strategies with use of stochastic dominance. In this case, there might be a chance that such stochastic dominance between two strategies’ utilities may not hold for any non-decreasing function  $U$ . Nevertheless, we believe that we can still compare the expected utilities, but it may require more sophisticated method to analyze on it.

## B Preliminaries to The Analysis

Our result on replication-proneness and replication-proofness can be shown in more generalized statements than the ones provided in the main paper. In this context, before presenting the main proofs, we provide a generalized model for our problem, then introduce some preliminaries required in presenting our results. We mainly consider two extensions: discount sequence and generalized utility function. Throughout the appendix, we abuse  $I_t(\mathcal{S}_i, \mathcal{S}_{-i})$  or  $I_t$  to denote the arm selected at round  $t$  instead of  $I_t(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T)$  if the context is clear.

### B.1 Discount sequence

As discussed, we consider a scenario that each agent  $i$  receives the sum of rewards from its arms. In practice, this corresponds to the scenario where a fixed portion of the rewards may be shared from the principal to the agents. Then, agent’s revenue can precisely be defined as  $v_i(\mathcal{S}_i, \mathcal{S}_{-i}) \cdot \varepsilon$ . However, we can ignore this constant since we’re only interested in comparing the agent’s utility between two or more strategies, where the impact of  $\varepsilon$  will be wiped out in this case. Hence, we re-define the following notion of agent’s revenue by introducing a discount sequence:

**Definition 5** (Agent’s discounted revenue). *Given a principal’s algorithm  $\mathfrak{A}$ , time horizon  $T \geq n$ , other agents’ strategic decision  $\mathcal{S}_{-i} = (\mathcal{S}_1, \dots, \mathcal{S}_{i-1}, \mathcal{S}_{i+1}, \dots, \mathcal{S}_n)$ , and discount factor  $\gamma = (\gamma_t)_{t=1}^T$ , agent  $i$ ’s revenue is defined as,*

$$v_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T) = \sum_{t=1}^T \gamma_t R_t .$$

The discount sequence captures the agent’s various viewpoint in computing long-term cumulative reward, e.g. agents’ utility can be discounted over time ( $\gamma_t = 1/t, t \in \mathbb{Z}_{\geq 0}$ ) or not ( $\gamma_t = 1, t \in \mathbb{Z}_{\geq 0}$ ). This generalized the practical scenario where the advertisers or content providers in online platform are usually myopic than the platform [Amin et al., 2013, Mohri and Medina, 2014]. We define a discount sequence to be *proper* if it is non-negative real-valued ( $\gamma_t \in \mathbb{R}_{\geq 0}$ ), non-increasing ( $\gamma_t \geq \gamma_{t+1}$ ), and has at least  $n$  non-zero elements. We denote the family of all proper discount sequences given  $T$  be  $\Gamma_T$ .

## B.2 Agent’s expected utility

Agent’s revenue is actually a random variable which can be varying with respect to the randomness of the policy, and the randomness of realized reward. Hence, there could be difference between the agents in how they assess the risk in the random revenue. For example, some agents might want to avoid a large variability in the random revenue, but some might love it. In this context, we consider the following general notion of agent’s expected utility, instead of agent’s expected revenue.

**Definition 6** (Agent’s expected utility). *Given a utility function  $U : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ , agent  $i$ ’s expected utility is defined as,*

$$u_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T) = \mathbb{E}[U(v_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T))].$$

Utility function  $U$  represents how the agents assess an uncertainty in revenue. If  $U$  is concave, they gradually underestimate the revenue growth as its absolute amount becomes large, hence in this case they prefer a certain amount of revenue rather than taking a risk to get larger revenue even if the expected revenue is smaller. We say that the agents are *risk-averse* in this case. In comparison, we call the agents be *risk-seeking* if  $U$  is convex, since they are willing to take a risk for having a large amount of revenue with small probability. If  $U$  is an identity function, we say that the agents are *risk-neutral* and they only care about the expected amount of revenue. In this case, expected utility simply reduces to the expected revenue which was provided in our main paper:

$$\sum_{t=1}^T \gamma_t \sum_{a \in \mathcal{S}_i} \mu(a) \cdot \mathbb{E}[\mathbb{1}[I_t(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T) = a]].$$

Although we consider a homogeneous scenario so that every agents share a common utility function, all the results can easily be extended to heterogeneous utility scenario.

## B.3 Further notions and lemmas

Before getting into the proofs, we introduce some notions to make the explanation more clear. We denote a pair of multi-armed bandit problem instance and corresponding algorithm be  $\mathcal{X} = (\langle T, \mathcal{S} \rangle, \mathfrak{A})$  for time horizon  $T$ , set of arms  $\mathcal{S}$  (with corresponding set of reward distributions  $d(\cdot)$ ), and MAB algorithm  $\mathfrak{A}$ . We call such  $\mathcal{X}$  be *MAB process*. We often omit  $T$  in  $\mathcal{X}$  as  $\mathcal{X} = (\langle \cdot, \mathcal{S} \rangle, \mathfrak{A})$  if  $\mathfrak{A}$  is horizon-independent. Given a MAB process  $\mathcal{X}$ , its sample path  $r_T^{\mathcal{X}} = (i_t, x_t)_{t \in [T]}$ , and any set of arms  $\mathcal{A} \subset \mathcal{S}$ , we define  $\mathcal{A}$ -sample path  $c_{\mathcal{A}}(r_T^{\mathcal{X}})$  be  $r_T^{\mathcal{X}}$ ’s longest subsequence that consists of elements contained in  $\mathcal{A}$ , i.e.

$$c_{\mathcal{A}}(r_T^{\mathcal{X}}) = \underset{\tau}{\operatorname{argmax}} \{ |r| : r = (i_{n_t}, x_{n_t})_{t \in [T]}, 1 \leq n_1 \leq n_2 \leq \dots \leq n_T \leq T, i_{n_t} \in \mathcal{A} \}.$$

We define  $t$ -count  $N_{\mathcal{A}, t}^{\mathcal{X}}$  be the number of times arm  $i \in \mathcal{A}$  is pulled in first  $t$  rounds in  $\mathcal{X}$ . We often omit  $\mathcal{X}$  in  $N_{\mathcal{A}, t}^{\mathcal{X}}$  if no confusion arises. We abuse the notation as  $N_{a, t}^{\mathcal{X}}$  when we consider a single arm  $a$  instead of a set of arms.

Next, we define a notion of stochastic dominance between two random variables as the following:

**Definition 7** (Stochastic dominance). *Let  $X$  and  $Y$  be two random variables.  $X$  has stochastic dominance over  $Y$  and we denote  $X \succeq Y$  if  $\forall x \in \mathbb{R}, \mathbb{P}[X \leq x] \leq \mathbb{P}[Y \leq x]$ . If  $X \succeq Y$  and there exists some  $x$  such that  $\mathbb{P}[X \leq x] < \mathbb{P}[Y \leq x]$ , then we say that  $X$  has strict stochastic dominance over  $Y$ ,  $X \succ Y$ .*

We note that stochastic dominance always induce a higher expectation since  $\mathbb{E}[X] = \sum_{x=0}^{\infty} \mathbb{P}[X > x] \geq \sum_{x=0}^{\infty} \mathbb{P}[Y > x] = \mathbb{E}[Y]$ , i.e. it is stronger ordering compared to ordering with expectation. Our definition of stochastic dominance refers to the first-order stochastic dominance in the literature, and it enables us to analyze the behavior of agents regardless of their behavior against the risk.

We now introduce some well-known properties on stochastic dominance which are useful in proving our main results. The following proposition provides a connection between stochastic dominance and expected utility:

**Proposition 1.** *Given random variables  $X$  and  $Y$ , the following holds:*

- 1)  $X \succeq Y$  if and only if  $\mathbb{E}[U(X)] \geq \mathbb{E}[U(Y)]$  for any non-decreasing utility function  $U(\cdot)$ ,

2)  $X \succ Y$  if and only if  $\mathbb{E}[U(X)] > \mathbb{E}[U(Y)]$  for any strictly increasing utility function  $U(\cdot)$ .

*Proof.* We only prove for the discrete case since it can easily be generalized. Suppose that  $X$  and  $Y$  are discrete random variables with outcome  $\{x_1, x_2, \dots, x_k\}$  where  $x_1 < x_2 < \dots < x_k$ .

Let  $d_i = \mathbb{P}[Y \leq x_i] - \mathbb{P}[X \leq x_i]$  and we have

$$\begin{aligned} \mathbb{E}[U(X)] - \mathbb{E}[U(Y)] &= \sum_{i=1}^k (\mathbb{P}[X = x_i] - \mathbb{P}[Y = x_i])U(x_i) \\ &= \sum_{i=1}^{k-1} (\mathbb{P}[X \leq x_i] - \mathbb{P}[Y \leq x_i])(U(x_i) - U(x_{i+1})) + (\mathbb{P}[X \leq x_k] - \mathbb{P}[Y \leq x_k])u_k \\ &= \sum_{i=1}^{k-1} (U(x_{i+1}) - U(x_i))d_i, \end{aligned}$$

where the second equation holds from Abel's lemma and the last comes from  $d_k = 0$ . It is straightforward that if  $d_i \geq 0$  for  $\forall i = 1, 2, \dots, k$  then  $\mathbb{E}[U(X)] \geq \mathbb{E}[U(Y)]$  since  $U$  is non-decreasing. Suppose that  $\mathbb{E}[U(X)] \geq \mathbb{E}[U(Y)]$  for any non-decreasing function  $U$  and there exists  $j$  such that  $d_j < 0$ . If we consider a utility function  $U$  such that  $U(x_{i+1}) = U(x_i)$  for  $i \neq j$ , and  $U(x_{j+1}) > U(x_j)$ , then  $\mathbb{E}[U(X)] - \mathbb{E}[U(Y)] = d_j(U(x_{j+1}) - U(x_j)) < 0$  which is a contradiction. We omit the proof for the second statement since it can easily be derived in similar argument.  $\square$

We also note that the stochastic dominance is preserved under summation and multiplication between positive independent random variables:

**Proposition 2.**  $X_1, X_2, Y_1$  and  $Y_2$  are positive random variables such that  $X_1$  and  $Y_1$  is independent, and  $X_2$  and  $Y_2$  is independent. If  $X_1 \succeq Y_1$  and  $X_2 \succeq Y_2$ , then, (i)  $X_1 + X_2 \succeq Y_1 + Y_2$  and (ii)  $X_1 X_2 \succeq Y_1 Y_2$ .

*Proof.* Since stochastic dominance is equivalent to the existence of monotone coupling [Roch, 2015], we can construct monotone couplings  $(\hat{X}_1, \hat{Y}_1)$  and  $(\hat{X}_2, \hat{Y}_2)$  of  $(X_1, Y_1)$  and  $(X_2, Y_2)$ , respectively. Then, we have  $X_1 + X_2 \sim \hat{X}_1 + \hat{X}_2 \succeq \hat{Y}_1 + \hat{Y}_2 \sim Y_1 + Y_2$  and  $X_1 X_2 \sim \hat{X}_1 \hat{X}_2 \succeq \hat{Y}_1 \hat{Y}_2 \sim Y_1 Y_2$  where  $\sim$  refers to the equality in distribution, and it concludes the proof.  $\square$

Finally, with use of stochastic dominance, the following theorem represents a sufficient condition for a strategy to be dominant over another strategy for any non-decreasing utility function  $U$ . We remark that we often say that a strategy is dominant over another strategy when it induces an expected utility which is at least that of another.

**Claim 1.** Given a principal's algorithm  $\mathfrak{A}$ , time horizon  $T \geq n$ , and other agents' strategic decision  $\mathcal{S}_{-i}$ , suppose that agent  $i$ 's two strategies  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  satisfy the followings:

- 1)  $a' \succeq a$  for  $\forall a' \in \mathcal{S}'_i$  and  $\forall a \in \mathcal{S}_i$ ,
- 2)  $N_{\mathcal{S}'_i, t}^{\mathcal{X}'}$  has stochastic dominance over  $N_{\mathcal{S}_i, t}^{\mathcal{X}}$  for any  $t \in [T]$

Then,  $\mathcal{S}'_i$  is a dominant over  $\mathcal{S}_i$  under any non-decreasing utility function  $U$  and any proper discount sequence  $\gamma \in \Gamma_T$ .

*Proof.* We denote  $r(a)$  be the revenue R.V. of arm  $a$ . The probability that agent  $i$ 's revenue is at most  $x$  can be

computed as the following:

$$\mathbb{P}_{\mathcal{X}}[v_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T) \leq x] = \mathbb{P}_{\mathcal{X}}\left[\sum_{t=1}^T \gamma_t r_{I_t(\mathcal{S}_i, \mathcal{S}_{-i}), t} \leq x\right] \quad (4)$$

$$= \mathbb{P}_{\mathcal{X}}\left[\sum_{t=1}^T \sum_{a \in \mathcal{S}_i} \gamma_t r(a) \mathbb{1}[I_t(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T) = a] \leq x\right] \quad (5)$$

$$= \mathbb{P}_{\mathcal{X}}\left[\left(\sum_{t=2}^T \gamma_t \sum_{a \in \mathcal{S}_i} r(a)(N_{a,t}^{\mathcal{X}} - N_{a,t-1}^{\mathcal{X}})\right) + \gamma_1 \sum_{a \in \mathcal{S}_i} \mu(a) N_{a,1}^{\mathcal{X}} \leq x\right] \quad (6)$$

$$= \mathbb{P}_{\mathcal{X}}\left[\sum_{t=1}^{T-1} \sum_{a \in \mathcal{S}_i} r(a)(\gamma_t - \gamma_{t+1}) N_{a,t}^{\mathcal{X}} \leq x\right] \quad (7)$$

$$\geq \mathbb{P}_{\mathcal{X}'}\left[\sum_{t=1}^{T-1} \sum_{a \in \mathcal{S}'_i} r(a)(\gamma_t - \gamma_{t+1}) N_{a,t}^{\mathcal{X}'} \leq x\right] = \mathbb{P}_{\mathcal{X}'}[v_i(\mathcal{S}'_i, \mathcal{S}_{-i}; \mathfrak{A}, T) \leq x], \quad (8)$$

where equation (6) holds since  $N_{a,t}^{\mathcal{X}} = N_{a,t-1}^{\mathcal{X}} + \mathbb{1}[I_t = a]$  for  $t \geq 2$  and  $N_{a,1}^{\mathcal{X}} = \mathbb{1}[I_1 = a]$ . In equation (7) we assume that  $\gamma_0 = 0$ , and inequality (8) follows from the stochastic dominance of  $N_{\mathcal{S}'_i, t}^{\mathcal{X}'}$  over  $N_{\mathcal{S}_i, t}^{\mathcal{X}}$  and condition 1 in theorem statement applied with 2. Hence we conclude that revenue under  $\mathcal{S}'_i$  has stochastic dominance over that under  $\mathcal{S}_i$ , and by Proposition 1,  $u_i(\mathcal{S}'_i, \mathcal{S}_{-i}; \mathfrak{A}, T) \geq u_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T)$  for any non-decreasing utility  $U$ .  $\square$

**Remark.** It is straightforward to check that  $u_i(\mathcal{S}'_i, \mathcal{S}_{-i}; \mathfrak{A}, T) > u_i(\mathcal{S}_i, \mathcal{S}_{-i}; \mathfrak{A}, T)$  if any of the described inequalities is strict. For example, if there exists some  $x$  and  $t$  such that  $\mathbb{P}[N_{a,t}^{\mathcal{X}} \leq x] < \mathbb{P}[N_{a,t}^{\mathcal{X}'} \leq x]$ , then the dominance is strict.

## B.4 Embedding

We now lay out the foundations in defining the embedding from a MAB process to the other MAB process which helps us to compare the agent's expected utility between two strategies. This will mainly be used at analyzing the replication-proneness. Given any two MAB processes  $\mathcal{X} = (\langle \cdot, \mathcal{S} \rangle, \mathfrak{A})$  and  $\mathcal{X}' = (\langle \cdot, \mathcal{S}' \rangle, \mathfrak{A}')$  with two horizon-free algorithms  $\mathfrak{A}$  and  $\mathfrak{A}'$ , we define an arm-translation  $\tau : \mathcal{S} \mapsto \mathcal{S}'$  be a mapping function between arms in MAB processes, and let  $Im_{\tau}(\mathcal{A})$  be image of  $\mathcal{A}$  under  $\tau$  for any  $\mathcal{A} \subset \mathcal{S}$ . Given any sample path  $r$ , let  $L(r)$  be its length, e.g. if  $r = (i_1, x_1) \times (i_2, x_2)$  then  $L(r) = 2$ , and  $\mathfrak{T}_{\tau}(r)$  be the sequence that replaces each selected arm  $a$  at round  $t$  in  $r$  into  $\tau(a)$  for any  $t \in [L(r)]$ . Now we define an *embedding* of MAB process  $\mathcal{X}$  into  $\mathcal{X}'$ .

**Definition 8** (Embedding). *Given an arm-translation  $\tau$ , we say that  $\mathcal{X}$  is embedded into  $\mathcal{X}'$  by  $\tau$  if a sample path  $r_t^{\mathcal{X}}$  is realized under  $\mathcal{X}$ , then sample path  $r_t^{\mathcal{X}'}$  is realized under  $\mathcal{X}'$  such that its  $\mathcal{S}_I$ -sample path is  $\mathfrak{T}_{\tau}(r_t^{\mathcal{X}})$ , i.e.  $\mathfrak{T}_{\tau}(r_t^{\mathcal{X}}) = c_{\mathcal{S}_I}(r_t^{\mathcal{X}'})$  for  $\mathcal{S}_I = Im_{\tau}(\mathcal{S})$ .*

Note that given  $\tau$ , we can construct an embedding of  $\mathcal{X}$  into  $\mathcal{X}'$  by coupling the sample paths in  $\mathcal{X}$  and  $\mathcal{X}'$  as the following: at each round in  $\mathcal{X}'$ , (i) arm  $a \in \mathcal{S}' \setminus \mathcal{S}_I$  can be selected in  $\mathcal{X}'$ , (ii)  $a \in \mathcal{S}_I$  can be selected in  $\mathcal{X}'$  and one of the arms in  $\tau^{-1}(a)$  can be selected in  $\mathcal{X}$ . In this scenario,  $\mathcal{X}$  move forwards only when  $a \in \mathcal{S}_I$  is selected under  $\mathcal{X}'$ . Likewise, we can enforce the described mapping between any realization of sample paths regardless of MAB processes  $\mathcal{X}$  and  $\mathcal{X}'$ , and hence we can always embed  $\mathcal{X}$  into  $\mathcal{X}'$ . However, it is not true that such mapped pair of sample paths has the same marginal probability to be realized in corresponding MAB process, hence we introduce the following notion of proper embedding:

**Definition 9** (Proper embedding). *We say that  $\mathcal{X}$  is properly embedded into  $\mathcal{X}'$  if for any  $t \geq 1$ , there exists a mapping  $\tau : \mathcal{S} \mapsto \mathcal{S}'$  such that given any sample path  $r_t^{\mathcal{X}}$  in  $\mathcal{X}$ , the marginal probability that sample path  $r$  is realized under  $\mathcal{X}$  whose  $\mathfrak{T}_{\tau}$  transformation is  $\mathfrak{T}_{\tau}(r_t^{\mathcal{X}})$ , is equal to the summation of marginal probability that sample path  $r'$  is realized of which its  $\mathcal{S}_I$ -sample path is  $\mathfrak{T}_{\tau}(r_t^{\mathcal{X}})$ , i.e.*

$$\mathbb{P}_{\mathcal{X}}[\{r : \mathfrak{T}_{\tau}(r) = \mathfrak{T}_{\tau}(r_t^{\mathcal{X}})\}] = \mathbb{P}_{\mathcal{X}'}[\{r' : c_{\mathcal{S}_I}(r') = \mathfrak{T}_{\tau}(r_t^{\mathcal{X}})\}].$$

Unlike from embedding, proper embedding does not always exist, but rather depends on the parameters of two MAB process.

**Illustrative examples.** To give an intuition on when the proper embedding exists and how it can be constructed, we provide some concrete examples.

**Example 1.** Suppose that we have two MAB process  $\mathcal{X} = (\langle \cdot, \{1, 2\} \rangle, \mathfrak{A})$  and  $\mathcal{X}' = (\langle \cdot, \{3, 4\} \rangle, \mathfrak{A}')$  where  $\mathfrak{A}$  selects the arms in uniform random manner, but  $\mathfrak{A}'$  selects arm 3 with probability  $1/3$  and 4 with probability  $2/3$ . If  $Im_\tau(\{1, 2\}) = \{3\}$  or  $\{4\}$ , then it is obvious that it is not proper embedding since both arms need to be selected in  $\mathcal{X}'$ . Suppose that  $\tau(1) = 3$  and  $\tau(2) = 4$  without loss of generality and sample path  $r = (1, \cdot) \times (2, \cdot)$  is given. In this case, the only sample path in  $\mathcal{X}'$  whose  $Im_\tau(\{1, 2\})$ -sample path is  $(1, \cdot) \times (2, \cdot)$ , is  $r' = (3, \cdot) \times (4, \cdot)$ . However,  $\mathbb{P}_{\mathcal{X}}[r] = 1/2 \cdot 1/2 \neq 1/3 \cdot 2/3 = \mathbb{P}_{\mathcal{X}'}[r']$  and hence there can be no proper embedding.

**Example 2.** Suppose that  $\mathcal{X} = (\langle \cdot, \{1, 2\} \rangle, \mathfrak{A})$  and  $\mathcal{X}' = (\langle \cdot, \{1, 2, 3\} \rangle, \mathfrak{A}')$  where both  $\mathfrak{A}$  and  $\mathfrak{A}'$  chooses the arms in uniformly random manner with respect to corresponding MAB instance. Consider  $\tau : \{1, 2\} \mapsto \{1, 2, 3\}$  such that  $\tau(1) = 1$  and  $\tau(2) = 2$ . In this case, given any realization  $r$  under  $\mathcal{X}$ , we can easily check that the summation of marginal probability that sample path  $r'$  under  $\mathcal{X}'$  whose  $\{1, 2\}$ -sample path is  $r$  is equal to the marginal probability of having  $r$  under  $\mathcal{X}$ .

**Example 3.** Suppose that  $\mathcal{X} = (\langle \cdot, \{1, 2, 3\} \rangle, \mathfrak{A})$  and  $\mathcal{X}' = (\langle \cdot, \{1, 2\} \rangle, \mathfrak{A}')$  where  $\mathfrak{A}$  chooses arm 1 and 2 with probability  $1/4$  and 3 with probability  $1/2$ , and  $\mathfrak{A}'$  chooses the arm in uniformly random manner. Consider  $\tau : \{1, 2, 3\} \mapsto \{1, 2\}$  such that  $\tau(1) = \tau(2) = 1$  and  $\tau(3) = 2$ . Since the summation of probability that arm 1 or 2 is selected in  $\mathcal{X}$  is equal to the probability that arm 1 is selected in  $\mathcal{X}'$ , it is straightforward that  $\tau$  makes a proper embedding from  $\mathcal{X}$  to  $\mathcal{X}'$ .

Proper embedding enables us to compare the probabilistic properties of random variables defined on two MAB process respectively. Representatively, we present the following claim which holds upon surjective arm-translation function:

**Claim 2.** *Given a MAB process  $\mathcal{X} = (\langle \cdot, \mathcal{S} \rangle, \mathfrak{A})$  and its proper embedding  $\mathcal{X}' = (\langle \cdot, \mathcal{S}' \rangle, \mathfrak{A}')$  with  $\tau : \mathcal{S} \mapsto \mathcal{S}'$ , if  $\tau$  is surjective and each mapped pair  $a$  and  $\tau(a)$  have the same reward distribution for any  $a \in \mathcal{S}$ , then the principal's regret until any round  $t$  has the same distribution between  $\mathcal{X}$  and  $\mathcal{X}'$ .*

We skip the proof since it is obvious from the definition of proper embedding.

If  $\tau$  is not surjective, then there will be unmapped arms in  $\mathcal{X}'$ , and this remaining arms will increase the revenue of the agent who owns these arms since they will be selected more times due to the existence of the remaining arms. This means that the agent's  $t$ -count can be stochastically ordered between two strategies, and we formally present this result as follows:

**Claim 3.** *Given a MAB process  $\mathcal{X} = (\langle \cdot, \mathcal{S} \rangle, \mathfrak{A})$  and its proper embedding  $\mathcal{X}' = (\langle \cdot, \mathcal{S}' \rangle, \mathfrak{A}')$  with  $\tau : \mathcal{S} \mapsto \mathcal{S}'$ , for any set of arms  $\mathcal{A} \subset \mathcal{S}$ ,  $N_{\mathcal{A}', t}^{\mathcal{X}'}$  has stochastic dominance over  $N_{\mathcal{A}, t}^{\mathcal{X}}$  for any  $t \geq 1$  where  $\mathcal{A}' = Im_\tau(\mathcal{A}) \cup (\mathcal{S}' \setminus Im_\tau(\mathcal{S}))$ .*

*Proof.* Given a sample path  $r_t^{\mathcal{X}}$  in  $\mathcal{X}$ , suppose that a sample path  $r'$  in  $\mathcal{X}'$  satisfies that  $c_{\mathcal{A}}(r') = Im_\tau(r_t^{\mathcal{X}})$ . Let  $\mathcal{A}_I = Im_\tau(\mathcal{A})$  and  $\mathcal{S}_I = Im_\tau(\mathcal{S})$ . Under the embedding, it is obvious that  $L(r') \geq t$  and  $N_{\mathcal{A}, t}^{\mathcal{X}}(r_t^{\mathcal{X}}) = N_{\mathcal{A}_I, L(r')}^{\mathcal{X}'}(r')$ . Suppose that arms in  $\mathcal{A}_I$  and  $\mathcal{S}_I \setminus \mathcal{A}_I$  appears  $x$  and  $y$  times respectively from round  $t + 1$  to  $L(r')$  in  $r'$  under  $\mathcal{X}'$ . Then the following inequalities hold:

$$N_{\mathcal{A}_I, L(r')}^{\mathcal{X}'}(r') = N_{\mathcal{A}_I, t}^{\mathcal{X}'}(r') + x \leq N_{\mathcal{A}_I, t}^{\mathcal{X}'}(r') + x + y = N_{\mathcal{A}', t}^{\mathcal{X}'}(r'),$$

where the last equation holds since arms in  $\mathcal{S}' \setminus \mathcal{S}_I$  appears exactly  $x + y$  times from round 1 to round  $t$  in  $r'$



under  $\mathcal{X}'$ . Hence, the following inequalities hold:

$$\begin{aligned}
 \mathbb{P}_{\mathcal{X}}[N_{\mathcal{A},t}^{\mathcal{X}} \leq m] &= \sum_{j=0}^m \mathbb{P}_{\mathcal{X}}[N_{\mathcal{A},t}^{\mathcal{X}} = j] = \sum_{j=0}^m \sum_{N_{\mathcal{A},t}^{\mathcal{X}}(r_t^{\mathcal{X}})=j} \mathbb{P}_{\mathcal{X}}[r_t^{\mathcal{X}}] \\
 &= \sum_{j=0}^m \sum_{N_{\mathcal{A},L}^{\mathcal{X}'}(r')=j} \mathbb{P}_{\mathcal{X}'}[r'] \\
 &= \sum_{r'} \mathbb{P}_{\mathcal{X}'}[N_{\mathcal{A},L}^{\mathcal{X}'}(r') \leq m] \\
 &\geq \sum_{r'} \mathbb{P}_{\mathcal{X}'}[N_{\mathcal{A},t}^{\mathcal{X}'}(r') \leq m] = \mathbb{P}_{\mathcal{X}'}[N_{\mathcal{A},t}^{\mathcal{X}'} \leq m],
 \end{aligned}$$

and we conclude that  $N_{\mathcal{A},t}^{\mathcal{X}'}$  has stochastic dominance over  $N_{\mathcal{A},t}^{\mathcal{X}}$ .  $\square$

## C Proof of Main Results

We now provide proofs for all the theorems presented in our main paper.

### C.1 Proof of Theorem 1

*Proof.* Instead of proving Theorem 1, we prove the following generalized statement:

**Proposition 3** (Restatement of theorem 1 in main paper). *UCB1 is replication-prone under any strictly increasing utility function  $U$ , and  $\forall \gamma \in \Gamma_T$ .*

We first consider a simple scenario when all the agents have single original arm. It suffices to show that replicating  $k$  times is a strictly dominant strategy over replicating  $k - 1$  times for  $k = 1, 2, \dots$ . Noting that the proof can easily be extended into the case where  $|\mathcal{S}_i| \geq 2$ , we only provide the proof for the case when  $|\mathcal{S}_i| = 1$ . Let  $\mathcal{S}_i = \{1\}$  and  $\mathcal{S}_{-i} = \{2, 3, \dots, k\}$  be the union of other agents' arms. Let  $\mathcal{S}'_i = \{1, 1'\}$  be agent  $i$ 's another strategy where  $1'$  is a replication of arm 1. Now we want to prove that agent  $i$ 's payoff is strictly dominant in  $\mathcal{S}'_i$  over  $\mathcal{S}_i$ . Let  $\mathcal{S} = \mathcal{S}_i \cup \mathcal{S}_{-i}$  and  $\mathcal{S}' = \mathcal{S}'_i \cup \mathcal{S}_{-i}$ . We denote MAB process for the non-replicated case be  $\mathcal{X} = (\langle \cdot, \mathcal{S}, \mathcal{D} \rangle, \mathfrak{A})$  and for the replicated case be  $\mathcal{X}' = (\langle \cdot, \mathcal{S}', \mathcal{D}' \rangle, \mathfrak{A}')$  for corresponding UCB1 algorithms  $\mathfrak{A}$  and  $\mathfrak{A}'$ . We now show that there exists a proper embedding from  $\mathcal{X}$  into  $\mathcal{X}'$ , and conclude the proof by calculating the expected payoff of agent  $i$  with the notion of  $t$ -count.

Firstly we prove that there exists a proper embedding from  $\mathcal{X}$  into  $\mathcal{X}'$ . Consider an arm-translation  $\tau : \mathcal{S} \mapsto \mathcal{S}'$  such that  $\tau(a) = a$  for any  $a \in \mathcal{S}$ . To show that there exists a proper embedding, since  $\{r_x : \mathfrak{T}_\tau(r_x) = \mathfrak{T}_\tau(r)\} = \{r\}$  by our construction of  $\tau$ , we need to show that for any valid sample path  $r$  in  $\mathcal{X}$ ,

$$\mathbb{P}_{\mathcal{X}}[r] = \sum_{c_{\mathcal{S}}(r') = \mathfrak{T}_\tau(r)} \mathbb{P}_{\mathcal{X}'}[r']. \quad (9)$$

Since  $\tau$  is identity function we simply use  $r$  instead of  $\mathfrak{T}_\tau(r)$ . We use the proof by induction with respect to the length of sample path in  $\mathcal{X}$ . Assume that the sample path  $r$  under  $\mathcal{X}$  is given and its length is 1, i.e.  $L(r) = 1$ . Let the arm selected in  $r$  be  $a \in \mathcal{S}$ . Then,

$$\begin{aligned}
 \sum_{c_{\mathcal{S}}(r')=r} \mathbb{P}_{\mathcal{X}'}[r'] &= \mathbb{P}_{\mathcal{X}'}[I_1 = a] + \mathbb{P}_{\mathcal{X}'}[\{I_1 \neq a\} \cap \{I_2 = a\}] \\
 &= 1/(k+1) + 1/(k+1) \cdot 1/k = \mathbb{P}_{\mathcal{X}}[I_1 = a] = \mathbb{P}_{\mathcal{X}}[r],
 \end{aligned}$$

and hence equation (9) holds for  $L(r) = 1$ . Now suppose that equation (9) holds for any sample path  $r$  with  $L(r) = m$ . Again, let  $r$  be the sample path of length  $m+1$  under  $\mathcal{X}$ , and  $r_m$  be its sub-sequence that only accounts for the first  $m$  elements in  $r$ , e.g. if  $r = (1, 0.1) \times \dots \times (m, 0.1) \times (m+1, 0.1)$  then  $r_m = (1, 0.1) \times \dots \times (m, 0.1)$ .

Let the arm selected in  $r$  at round  $m + 1$  be  $a \in \mathcal{S}$ . Then the following equations hold:

$$\mathbb{P}_{\mathcal{X}}[r] = \mathbb{P}_{\mathcal{X}}[r|r_m]\mathbb{P}_{\mathcal{X}}[r_m] = \mathbb{P}_{\mathcal{X}}[I_{m+1} = a|r_m]\mathbb{P}_{\mathcal{X}}[r_m] \quad (10)$$

$$= \sum_{c_{\mathcal{S}}(r')=r_m} \mathbb{P}_{\mathcal{X}'}[r'] \cdot \mathbb{P}_{\mathcal{X}}[I_{m+1} = a|r_m]. \quad (11)$$

Now we look into  $\mathbb{P}_{\mathcal{X}}[I_{m+1} = a|r_m]$ . For  $s \in \mathcal{S}$ , let arm  $s$ 's UCB index at round  $m$  be  $u_m(s)$ . If  $u_m(a) < \operatorname{argmax}_{s \in \mathcal{S}}(u_m(s))$ , then arm  $a$  cannot be selected at round  $m + 1$  and hence  $r$  can't be realized under  $\mathcal{X}$ , which is a contradiction. Hence we have  $u_m(a) \geq \operatorname{argmax}_{s \in \mathcal{S}}(u_m(s))$ , and let the set of arms in this tied index be  $\mathcal{T}$ . Given  $r'$  with  $c_{\mathcal{S}}(r') = r_m$ , suppose that arm  $1'$  has the highest UCB index from round  $L(r)$  to  $L(r) + t - 1$ , and hence selected by UCB1 at rounds  $L(r) + 1$  to  $L(r) + t$ , and not for round  $L(r) + t$ . It is obvious that such  $t < \infty$  exists since arm  $1'$  and  $1$  follows the same reward distribution. In this case at round  $L(r) + t + 1$ , arm  $1'$  cannot have the highest UCB index, and hence one of the arms in  $\mathcal{T}$  would be selected since  $c_{\mathcal{S}}(r') = r_m$  and only arm  $1'$  is selected from round  $L(r) + 1$  to  $L(r) + t$ . Hence, the probability that arm  $a$  is selected in this case is equal to  $\mathbb{P}_{\mathcal{X}}[I_{m+1} = a|r_m]$ , and we have:

$$\sum_{c_{\mathcal{S}}(r')=r_m} \mathbb{P}_{\mathcal{X}'}[r'] \cdot \mathbb{P}_{\mathcal{X}}[I_{m+1} = a|r_m] = \sum_{c_{\mathcal{S}}(r')=r_m} \mathbb{P}_{\mathcal{X}'}[r']/|\mathcal{T}| \quad (12)$$

$$= \sum_{c_{\mathcal{S}}(r')=r_m} \mathbb{P}_{\mathcal{X}'}[r'] \cdot \mathbb{P}_{\mathcal{X}'}[I_{L(r)+t+1} = a|r'] \quad (13)$$

$$= \sum_{c_{\mathcal{S}}(r')=r} \mathbb{P}_{\mathcal{X}'}[r']. \quad (14)$$

By equation (11) and (14), we have  $\mathbb{P}_{\mathcal{X}}[r] = \sum_{c_{\mathcal{S}}(r')=r} \mathbb{P}_{\mathcal{X}'}[r']$  and we conclude that equation (9) is true. Hence, by Claim 1 and 3,  $\mathcal{S}'_i$  is a dominant strategy over  $\mathcal{S}_i$  regardless of the MAB instance, and now it is enough to prove that there exists a MAB instance such that equality does not hold in  $u_i(\mathcal{S}'_i, \mathcal{S}_{-i}) \geq u_i(\mathcal{S}_i, \mathcal{S}_{-i})$ . In Bernoulli bandit case, it is obvious that  $\mathbb{P}_{\mathcal{X}'}[N_{\mathcal{S}'_i, t}^{\mathcal{X}'} \geq 2] = 1 > \mathbb{P}_{\mathcal{X}}[N_{\mathcal{S}_i, t}^{\mathcal{X}} \geq 2]$  for any  $t \geq n$ . Then in inequality (8) in the proof of Claim 1, the equality cannot hold and hence  $\mathcal{S}'_i$  is a strict dominant strategy over  $\mathcal{S}_i$ , which implies that  $u_i(\mathcal{S}'_i, \mathcal{S}_{-i}) > u_i(\mathcal{S}_i, \mathcal{S}_{-i})$  for any time horizon  $T$  by Proposition 1.

Finally, we can extend our proof into the multiple arm cases by constructing  $\mathcal{S}'$  which replicates all the arms in  $\mathcal{S}_i$  once. We omit the detailed process since it can easily be derived.  $\square$

## C.2 Proof of Theorem 2

*Proof.* We prove the following generalized statement:

**Proposition 4** (Restatement of theorem 2 in main paper). *H-UCB is replication-proof for  $i \in \mathcal{N}$  under any non-decreasing utility function  $U$ ,  $\forall T \geq n$  and  $\forall \gamma \in \Gamma_T$ .*

Let  $\mathcal{X} = (\langle \cdot, \mathcal{S}_i \cup \mathcal{S}_{-i} \rangle, \mathfrak{A})$  and  $\mathcal{X}' = (\langle \cdot, \mathcal{S}'_i \cup \mathcal{S}_{-i} \rangle, \mathfrak{A})$  where  $\mathfrak{A}$  is H-UCB. By Claim 1, it is enough to show that  $N_{\mathcal{S}'_i, t}^{\mathcal{X}'}$  has stochastic dominance over  $N_{\mathcal{S}_i, t}^{\mathcal{X}}$  for any strategy  $\mathcal{S}'_i$ . To this end, we show that  $\mathbb{1}[I_t^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] \geq \mathbb{1}[I_t^{\mathcal{X}'}(\mathcal{S}'_i, \mathcal{S}_{-i}) \in \mathcal{S}'_i]$  for any round  $t$ . We use proof by induction. For  $t \in [n]$ , it is straightforward that  $\mathbb{1}[I_t^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] \sim \mathbb{1}[I_t^{\mathcal{X}'}(\mathcal{S}'_i, \mathcal{S}_{-i}) \in \mathcal{S}'_i]$  due to the agent initialization step in H-UCB. Suppose that we have  $\mathbb{1}[I_t^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] \geq \mathbb{1}[I_t^{\mathcal{X}'}(\mathcal{S}'_i, \mathcal{S}_{-i}) \in \mathcal{S}'_i]$  holds for any  $t \in [t']$ . Then, by Proposition 2, we can derive the following relation between  $t$ -count of  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  for any  $t \in [t']$ :

$$N_{\mathcal{S}_i, t}^{\mathcal{X}} = \sum_{q=1}^t \mathbb{1}[I_q^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] \geq \sum_{q=1}^t \mathbb{1}[I_q^{\mathcal{X}'}(\mathcal{S}'_i, \mathcal{S}_{-i}) \in \mathcal{S}'_i] = N_{\mathcal{S}'_i, t}^{\mathcal{X}'}.$$

The probability that agent  $i$  is selected at round  $t' + 1$  can be computed as the following<sup>4</sup>:

$$\begin{aligned} \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] &= \sum_{m=0}^{t'} \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m] \mathbb{P}[N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m] \\ &= \sum_{m=0}^{t'} \sum_{r \geq 0} \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m, \bar{r}_{i,m}^{\mathcal{X}} = r] \mathbb{P}[\bar{r}_{i,m}^{\mathcal{X}} = r | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m] \mathbb{P}[N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m], \end{aligned} \quad (15)$$

where  $\bar{r}_{i,m}^{\mathcal{X}}$  denotes empirical average reward of agent  $i$  given that agent  $i$  is selected  $m$ th time by H-UCB under MAB process  $\mathcal{X}$ . Since  $X_{i,1} \succeq X_{i,k}$  for any  $k \in [l(i)]$ , it is straightforward that  $\bar{r}_{i,m}^{\mathcal{X}}$  conditioned to  $N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m$  has stochastic dominance over  $\bar{r}_{i,m}^{\mathcal{X}'}$  conditioned to  $N_{\mathcal{S}_i, t'}^{\mathcal{X}'} = m$ , i.e.

$$\mathbb{P}[\bar{r}_{i,m}^{\mathcal{X}} > r | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m] \geq \mathbb{P}[\bar{r}_{i,m}^{\mathcal{X}'} > r | N_{\mathcal{S}_i, t'}^{\mathcal{X}'} = m], \quad (16)$$

for any  $r \geq 0$ .

Besides, due to the nature of phase 1 that it selects agent solely based on agent's  $t$ -count and empirical average reward, the following holds for any  $r_1 \geq r_2$ :

$$\mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m, \bar{r}_{i,m} = r_1] \geq \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m, \bar{r}_{i,m}^{\mathcal{X}} = r_2]. \quad (17)$$

Now we observe that the distribution of sample path for the arms in  $\mathcal{S}_i'$  conditioned on agent  $i$ 's  $t$ -count and empirical average reward would be the same<sup>5</sup> between  $\mathcal{X}$  and  $\mathcal{X}'$ . This implies that the probability that agent  $i$  is selected at round  $t$  given its  $t$ -count and empirical average reward is the same between  $\mathcal{X}$  and  $\mathcal{X}'$ , and hence equation (17) can be lower-bounded as the following for any  $r_1 \geq r_2$ :

$$\mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m, \bar{r}_{i,m} = r_1] \geq \mathbb{P}[I_{t'+1}^{\mathcal{X}'}(\mathcal{S}_i', \mathcal{S}_{-i}') \in \mathcal{S}_i' | N_{\mathcal{S}_i', t'}^{\mathcal{X}'} = m, \bar{r}_{i,m}^{\mathcal{X}'} = r_2]. \quad (18)$$

Since we have  $N_{\mathcal{S}_i, t}^{\mathcal{X}} \succeq N_{\mathcal{S}_i', t'}^{\mathcal{X}'}$  for any  $t \in [t']$  by our induction hypothesis, plugging inequalities (16) and (18) into (15) yields the following:

$$\begin{aligned} \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] &= \sum_{m=0}^t \sum_{r \geq 0} \mathbb{P}[I_{t'+1}^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m, \bar{r}_{i,m}^{\mathcal{X}} = r] \mathbb{P}[\bar{r}_{i,m}^{\mathcal{X}} = r | N_{\mathcal{S}_i, t'}^{\mathcal{X}} = m] \mathbb{P}[N_{\mathcal{S}_i, t}^{\mathcal{X}} = m] \\ &\geq \sum_{m=0}^t \sum_r \mathbb{P}[I_{t'+1}^{\mathcal{X}'}(\mathcal{S}_i', \mathcal{S}_{-i}') \in \mathcal{S}_i' | N_{\mathcal{S}_i', t'}^{\mathcal{X}'} = m, \bar{r}_{i,m} = r] \mathbb{P}[\bar{r}_{i,m} = r | N_{\mathcal{S}_i', t'}^{\mathcal{X}'} = m] \mathbb{P}[N_{\mathcal{S}_i', t'}^{\mathcal{X}'} = m] \\ &= \mathbb{P}[I_{t+1}^{\mathcal{X}'}(\mathcal{S}_i', \mathcal{S}_{-i}') \in \mathcal{S}_i']. \end{aligned}$$

Hence we conclude that  $\mathbb{1}[I_t^{\mathcal{X}}(\mathcal{S}_i, \mathcal{S}_{-i}) \in \mathcal{S}_i] \succeq \mathbb{1}[I_t^{\mathcal{X}'}(\mathcal{S}_i', \mathcal{S}_{-i}') \in \mathcal{S}_i']$  for any  $t \geq 1$ , which implies that  $N_{\mathcal{S}_i, t}^{\mathcal{X}} \succeq N_{\mathcal{S}_i', t}^{\mathcal{X}'}$  for any  $t \geq 1$  by Proposition 2. By Claim 1, we conclude that H-UCB is replication-proof.  $\square$

### C.3 Proof of Theorem 3

Firstly, under the equilibrium  $\mathcal{S} = \cup_{i \in \mathcal{N}} \mathcal{S}_i$  such that  $\mathcal{S}_i = \{o_{i,1}\}$  for any  $i \in \mathcal{N}$ , it is straightforward to check that H-UCB submits the proposed regret bound since it exactly reduces to UCB1 under the canonical setting. Now we characterize all the possible dominant strategy equilibrium, and then show that there exists a proper embedding for any pair of equilibrium. Suppose that  $\mathcal{S}_i'$  possesses a suboptimal arm  $a$ . Then, inequality (16) in the proof of Theorem 2 will obviously become strict. Following the similar analysis, we can observe that the

<sup>4</sup>In precise, we need to integrate over  $r \geq 0$  rather than summing over it since the support of  $\bar{r}_{i,m}^{\mathcal{X}}$  is continuous over non-negative real value, however, we abuse it for the sake of simplicity.

<sup>5</sup>One may consider a sort of partial proper embedding from  $\mathcal{X}$  to  $\mathcal{X}'$  so that any arm  $a \in \mathcal{S}_i'$  in  $\mathcal{X}$  is mapped to  $a \in \mathcal{S}_i'$  in  $\mathcal{X}'$ , which implies that the selection and realization of the arms in  $\mathcal{S}_i'$  can be coupled between  $\mathcal{X}$  and  $\mathcal{X}'$ .

selection probability of agent  $i$  under  $\mathcal{S}'_i$  will be strictly smaller than that under  $\mathcal{S}_i$ , which eventually implies that  $\mathcal{S}'_i$  cannot be a dominant strategy. Finally, we prove that the principal's regret remains the same as  $\mathcal{S}_i$  under any strategy  $\mathcal{S}''_i = \{o_{i,1}^{(1)}, o_{i,1}^{(2)}, \dots, o_{i,1}^{(c_{i,1})}\}$ . Let  $\mathcal{X}'' = (\langle \cdot, \mathcal{S}''_i, \mathcal{S}_{-i} \rangle, \mathfrak{A})$ , and  $\mathcal{S}'' = \mathcal{S}''_i \cup \mathcal{S}_{-i}$ . Consider arm-translation  $\tau : \mathcal{S}'' \rightarrow \mathcal{S}$  which maps all the arms in  $\mathcal{S}''_i$  to  $o_{i,1}$  in  $\mathcal{S}_i$  and identically maps the arms in  $\mathcal{S}_{-i}$ . Then, it is straightforward to check that this constructs a proper embedding from  $\mathcal{X}''$  to  $\mathcal{X}$ . Hence by Claim 2, the principal's regret remains the same under  $\mathcal{S}$  and  $\mathcal{S}''$ , and we conclude the proof.

#### C.4 Proof of Theorem 5

*Proof.* For simplicity, we assume that  $\mathcal{S}_i$  is ordered so that  $\mu(s_{i,1}) \geq \mu(s_{i,2}) \geq \dots \geq \mu(s_{i,l(i)})$  for  $i \in \mathcal{N}$ . For simplicity, we abuse  $N_{\mathcal{A},t}$  to denote  $N_{\mathcal{A},t}^{\mathcal{X}}$ . We use  $N_{i,t}$ ,  $N_{i_a,t}$  to denote the expected number of times agent  $i$  and arm  $i_a$  are played up to time  $t$ , and  $R_{i,t}$ ,  $R_{i_a,t}$  be the empirical average reward of agent  $i$  and arm  $i_a$  at round  $t$ , respectively. Hence, if the index is contained in  $\mathcal{N}$ , it refers to the agent's random variable, otherwise, it indicates to that of the arm. Let  $i^*$  be the agent with optimal arm, and  $a^*$  be the optimal arm. Let  $\Delta_i = \mu(s_{i^*,1}) - \mu(s_{i,1})$ ,  $\Delta_{i_a} = \mu(s_{i^*,1}) - \mu(i_a)$ ,  $\Delta_i^m = \mu(s_{i^*,1}) - \mu(s_{i,l(i)})$ , and  $\delta_{i_a} = \mu(s_{i,1}) - \mu(i_a)$ .

Thanks to our analysis given in so far, any strategic agent can be viewed as a replicator with only single original arm, since it will only register the best arm or any set of replicas of it. Hence we assume that there only exists replicators without loss of generality.<sup>6</sup>

Let  $Y_i$  be the event that  $\mathcal{B}_i$  contains  $i$ 's optimal arm or any replica of it. We begin with the following two lemmas which gives upper-bound for expected regret occurred from replicator  $i$ .

**Lemma 1.**  $\mathbb{P}[Y_i^c] \leq 1/T$  .

*Proof of Lemma 1.* If  $|\mathcal{S}_i| \leq L \ln T$ , then  $\mathbb{P}[Y_i] = 1$ . If  $\mathcal{S}_i$ 's cardinality is infinite, then  $\mathbb{P}[Y_i^c] = (1-c)^{L \ln T}$ . Otherwise, the probability that  $i$ 's optimal arm is not sampled  $\mathcal{B}_i$  is given by

$$\begin{aligned} \mathbb{P}[Y_i^c] &\leq (1-c) \frac{|\mathcal{S}_i|(1-c) - 1}{|\mathcal{S}_i| - 1} \frac{|\mathcal{S}_i|(1-c) - 2}{|\mathcal{S}_i| - 2} \cdots \frac{|\mathcal{S}_i|(1-c) - L \ln T}{|\mathcal{S}_i| - L \ln T} \\ &\leq (1-c)^{L \ln T}. \end{aligned}$$

Since  $(1 - 1/x)^x \leq 1/e$  for any  $x \geq 0$ , we can derive the following bound on  $(1-c)^{L \ln T}$ :

$$(1-c)^{L \ln T} \leq (1-c)^{1/c \ln T} \leq \left(\frac{1}{e}\right)^{\ln T} = 1/T .$$

□

**Lemma 2.** *Given that event  $Y_i$  occurs, its conditional expected internal regret is bounded as the following:*

$$\sum_{i_a \in \mathcal{B}_i} \mathbb{E}[\delta_{i_a} N_{i_a,t} | Y_i] \leq L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{8 \ln N_{i,t}}{\delta_{i_a}} + \left(1 + \frac{\pi^2}{3}\right) \delta_{i_a} \right) .$$

*Proof of Lemma 2.* If the optimal arm is sampled at least one time, then all the other sampled arms induce  $O(\ln T)$  internal regret [Auer et al., 2002] since the behavior of phase 2 in RH-UCB given the sampled arms is equal to that of UCB1 under the canonical MAB problem. Hence, we have

$$\sum_{i_a \in \mathcal{B}_i} \mathbb{E}[\delta_{i_a} N_{i_a,t} | Y_i] \leq L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{8 \ln N_{i,t}}{\delta_{i_a}} + \left(1 + \frac{\pi^2}{3}\right) \delta_{i_a} \right) ,$$

from the regret analysis of [Auer et al., 2002]

□

Now, we prove the following bounds on the empirical average reward of any agent.

<sup>6</sup>Though, we can tighten our analysis by separating the regret from each type since the regret bound for strategic agent would definitely be smaller.

**Lemma 3.** *Given any agent  $i \in \mathcal{N}$ , for each arm  $j$  in  $j \in \mathcal{S}_i$  at time  $t$ , we have*

$$\mathbb{P}\left[|R_{i,t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}\right] \leq 1/T + 2L \ln T \exp\left(\frac{-\sqrt{t} \ln t}{2L^2 \ln^2 T}\right) + \frac{L \ln T}{\sqrt{N_{i,t} \sqrt{t} \ln t}} \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i,t}}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3}) \delta_{i_a} \right).$$

*Proof of Lemma 3.* We can divide the probability in our lemma as the following:

$$\mathbb{P}\left[|R_{i,t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}\right] = \mathbb{P}\left[Y_i \cap |R_{i,t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}\right] + \mathbb{P}\left[Y_i^c \cap |R_{i,t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}\right]. \quad (19)$$

The latter term is at most  $1/T$  by Lemma 1. For the former one, we begin with the following inequalities where we use  $\mathbb{P}_{Y_i}[X] = \mathbb{P}[X|Y_i]$  for simplicity:

$$\begin{aligned} & \mathbb{P}\left[Y_i \cap |R_{i,t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}\right] \\ & \leq \mathbb{P}_{Y_i}\left[|N_{i,t} R_{i,t} - N_{i,t} \mu_{i_1}| \geq \sqrt{N_{i,t} \sqrt{t} \ln t}\right] \\ & = \mathbb{P}_{Y_i}\left[\left|\sum_{a=1}^{\mathcal{S}_i} N_{i_a,t} R_{i_a,t} - N_{i,t} \mu_{i_1}\right| \geq \sqrt{N_{i,t} \sqrt{t} \ln t} \mid Y_i\right] \\ & = \mathbb{P}_{Y_i}\left[\left|\sum_{i_a \in \mathcal{B}_i} N_{i_a,t} R_{i_a,t} - \sum_{i_a \in \mathcal{B}_i} N_{i_a,t} \mu_{i_a} + \sum_{i_a \in \mathcal{B}_i} N_{i_a,t} \mu_{i_a} - \mu_{i_1} \sum_{i_a \in \mathcal{B}_i} N_{i_a,t}\right| \geq \sqrt{N_{i,t} \sqrt{t} \ln t}\right] \\ & \leq \mathbb{P}_{Y_i}\left[\left|\sum_{i_a \in \mathcal{B}_i} N_{i_a,t} (R_{i_a,t} - \mu_{i_a})\right| \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4}}\right] + \mathbb{P}_{Y_i}\left[\left|\sum_{i_a \in \mathcal{B}_i} N_{i_a,t} (\mu_{i_1} - \mu_{i_a})\right| \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4}}\right] \end{aligned} \quad (20)$$

$$\leq \mathbb{P}_{Y_i}\left[\sum_{i_a \in \mathcal{B}_i} |N_{i_a,t} (R_{i_a,t} - \mu_{i_a})| \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4}}\right] + \mathbb{P}_{Y_i}\left[\sum_{i_a \in \mathcal{B}_i} N_{i_a,t} (\mu_{i_1} - \mu_{i_a}) \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4}}\right] \quad (21)$$

$$\leq \sum_{i_a \in \mathcal{B}_i} \mathbb{P}_{Y_i}\left[|(R_{i_a,t} - \mu_{i_a})| \frac{N_{i_a,t}}{N_{i,t}} \geq \sqrt{\frac{\sqrt{t} \ln t}{4|\mathcal{B}_i^2|N_{i,t}}}\right] + \mathbb{P}_{Y_i}\left[\sum_{\substack{i_a \in \mathcal{B}_i \\ \mu_{i_a} < \mu_{i_1}}} \delta_{i_a} N_{i_a,t} \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4}}\right], \quad (22)$$

where we use  $\mathbb{P}(\sum_i^n X_i \geq x) \leq \sum_i^n \mathbb{P}(X_i \geq x/n)$  for inequality (20) and (22), and inequality (21) holds by triangle inequality. The first term can be bounded by Chernoff-Hoeffding inequality as follow:

$$\begin{aligned} \mathbb{P}_{Y_i}\left[|R_{i_a,t} - \mu_{i_a}| \cdot \frac{N_{i_a,t}}{N_{i,t}} \geq \sqrt{\frac{\sqrt{t} \ln t}{4|\mathcal{B}_i^2|N_{i,t}}}\right] & \leq \mathbb{P}_{Y_i}\left[|R_{i_a,t} - \mu_{i_a}| \cdot N_{i_a,t} \geq \sqrt{\frac{N_{i,t} \sqrt{t} \ln t}{4L^2 \ln^2 T}}\right] \leq 2 \exp\left(-\frac{\sqrt{t} \ln t N_{i,t}}{2L^2 \ln^2 T N_{i_a,t}}\right) \\ & \leq 2 \exp\left(-\frac{\sqrt{t} \ln t}{2L^2 \ln^2 T}\right). \end{aligned} \quad (23)$$

For the second term in inequality (22), the following holds:

$$\mathbb{P}_{Y_i}\left[\sum_{\substack{a \in \mathcal{B}_i \\ \mu_{i_a} < \mu_{i_1}}} \delta_{i_a} N_{i_a,t} \geq \sqrt{N_{i,t} \sqrt{t} \ln t / 4}\right] \leq 2\mathbb{E}\left[\sum_{\substack{a \in \mathcal{B}_i \\ \mu_{i_a} < \mu_{i_1}}} \delta_{i_a} N_{i_a,t} \mid Y_i\right] / \sqrt{N_{i,t} \sqrt{t} \ln t} \quad (24)$$

$$\leq L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i,t}}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3}) \delta_{i_a} \right) / \sqrt{N_{i,t} \sqrt{t} \ln t}, \quad (25)$$

where inequality (24) holds by conditional Markov inequality, and (25) follows from Lemma 2. Combining inequalities (19),(22),(23), and (25), we conclude the proof.  $\square$

In overall, our total regret can be divided into the regret (i) that internally incurs from any optimal agent<sup>7</sup> during identifying his optimal arm, and (ii) that externally incurs from selecting any suboptimal agents. We can simply bound the former one as the following, which is  $O(\ln^2 T)$ :

$$\mathbb{E}[\text{Regret}_T^{\text{opt}}] \leq T \cdot 1/T + L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln T}{\delta_{i_a}} + 9\delta_{i_a} \right), \quad (26)$$

by the fact that the expected regret given  $Y_i^c$  is at most  $T$ ,  $\mathbb{P}[Y_i^c] \leq 1/T$ , and Lemma 2. Hence we focus on the regret from any suboptimal agent.

Using Lemma 3, we upper bound the probability that suboptimal replicator  $i$  is selected given that replicator  $i$  and optimal replicator  $i^*$  has already selected at least certain times. For notational simplicity, we denote the bounding probability (terms in RHS) in Lemma 3 be  $\rho_{i,t}(N_{i,t})$ , i.e.

$$\rho_{i,t}(N_{i,t}) = 1/T + 2L \ln T e^{-\sqrt{t} \ln t / (2L^2 \ln^2 T)} + L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i,t}}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3})\delta_{i_a} \right) / \sqrt{N_{i,t} \sqrt{t} \ln t}.$$

We abuse  $I_t$  to denote the arm selected at round  $t$  if the context is clear. We define  $UCB_{i,t}$  to denote the UCB score of agent  $i$  at round  $t$ , i.e.  $UCB_{i,t} = R_{i,t} + \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}$ . The following lemma upper-bound the probability that any suboptimal replicator  $i$  is selected given that replicator  $i$  and any optimal replicator  $i^*$  is sufficiently selected:

**Lemma 4.** *At round  $t$ , if any suboptimal replicator  $i$  has been played at least  $\frac{4\sqrt{t} \ln t}{\Delta_i^2}$  times and optimal replicator  $i^*$  has been played at least  $\sqrt{t} \ln t$ , then  $UCB_{i,t} < UCB_{i^*,t}$  with probability at least  $1 - (\rho_{i,t}(\frac{4\sqrt{t} \ln t}{\Delta_i^2}) + \rho_{i^*,t}(\sqrt{t} \ln t))$ , i.e.*

$$\mathbb{P}\left[I_{t+1} \in \mathcal{S}_i \mid N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} \geq \sqrt{t} \ln t\right] \leq \rho_{i,t}\left(\frac{4\sqrt{t} \ln t}{\Delta_i^2}\right) + \rho_{i^*,t}(\sqrt{t} \ln t) \quad (27)$$

*Proof of Lemma 4.* Given  $N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}$  and  $N_{i^*,t} \geq \sqrt{t} \ln t$ , assume that event  $|R_{i,t} - \mu_{i_1}| \leq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}$  occurs for replicator  $i$  and  $i^*$ . Then we have

$$UCB_{i,t} = R_{i,t} + \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}} \leq R_{i,t} + \frac{\Delta_i}{2} \leq \mu_{i_1} + \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}} + \frac{\Delta_i}{2} \leq (\mu_{i_1} + \frac{\Delta_i}{2}) + \frac{\Delta_i}{2} = \mu^* < UCB_{i^*,t},$$

where the first inequality holds by  $N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}$ , the second holds by  $|R_{i,t} - \mu_{i_1}| \leq \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}$ , the third again holds by  $N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}$ , and the last follows from  $|R_{i^*,t} - \mu_{i_1^*}| \leq \sqrt{\frac{\sqrt{t} \ln t}{N_{i^*,t}}}$ .

By union bound, the probability of  $|R_{i,t} - \mu_{i_1}| > \sqrt{\frac{\sqrt{t} \ln t}{N_{i,t}}}$  or  $|R_{i^*,t} - \mu_{i_1^*}| > \sqrt{\frac{\sqrt{t} \ln t}{N_{i^*,t}}}$  is at most  $\rho_{i,t}(N_{i,t}) + \rho_{i^*,t}(N_{i^*,t})$ . Since  $\rho_{i,t}(\sqrt{t} \ln t)$  is a decreasing function on  $t$  for  $t \geq 1$ , we conclude the proof.  $\square$

Now for any suboptimal agent  $i$ , we upper bound the expected number of times it is played up to round  $T$  under

<sup>7</sup>In here, optimal agent refers to the agent with at least one optimal arm.

RH-UCB as the following:

$$\begin{aligned}
 \mathbb{E}[N_{i,T}] &= 1 + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i)\right] \\
 &= 1 + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i, N_{i,t} < \frac{4\sqrt{t} \ln t}{\Delta_i^2})\right] + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i, N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2})\right] \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i, N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2})\right] \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_{i^*}, N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} < \sqrt{t} \ln t)\right] \\
 &\quad + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i, N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} \geq \sqrt{t} \ln t)\right] \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + \mathbb{E}\left[\sum_{t=n}^T \mathbf{1}(I_{t+1} \in \mathcal{S}_i, N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} \geq \sqrt{t} \ln t)\right] \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + \sum_{t=n}^T \mathbb{P}\left[(I_{t+1} \in \mathcal{S}_i | N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} \geq \sqrt{t} \ln t)\right] \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + \sum_{t=n}^T \left(\rho_{i,t}\left(\frac{4\sqrt{t} \ln t}{\Delta_i^2}\right) + \rho_{i^*,t}(\sqrt{t} \ln t)\right), \tag{28}
 \end{aligned}$$

where inequality (28) holds since if  $N_{i,t} \geq 4\sqrt{t} \ln t / \Delta_i^2$  and  $N_{i^*,t} < \sqrt{t} \ln t / 4$  then  $R_{i,t} + \sqrt{\sqrt{t} \ln t / N_{i,t}} < 2 \leq \mathbb{R}_{i^*,t} + \sqrt{\sqrt{t} \ln t / N_{i^*,t}}$  and inequality (29) holds by Lemma 4.

By integral test, we have:

$$\sum_{t=2}^T \frac{1}{\sqrt{Ct} \ln t} \leq \sum_{t=2}^T \sqrt{\frac{1}{Ct}} \leq \frac{2}{\sqrt{C}} \int_{t=2}^{T+1} \frac{d\sqrt{t}}{dt} dt \leq 2\sqrt{T/C} \tag{30}$$

$$\sum_{t=2}^T 2L \ln T e^{-\frac{\sqrt{t} \ln t}{2L^2 \ln^2 T}} \leq \sum_{t=2}^T 2L \ln T e^{-\frac{\sqrt{t}}{2L^2 \ln^2 T}} \leq \int_{t=0}^{\infty} 4L^4 \ln^4 T \frac{d(-2e^{-\sqrt{t}}(\sqrt{t}+1))}{dt} = 8L^5 \ln^5 T. \tag{31}$$

Applying inequalities (30) and (31), we can bound the summation of  $\rho(C\sqrt{t} \ln t)$  as the following:

$$\begin{aligned}
 \sum_{t=n}^T \rho_{i,t}(C\sqrt{t} \ln t) &= \sum_{t=2}^T \left[1/T + 2L \ln T e^{-\sqrt{t} \ln t / (2L^2 \ln^2 T)} + L \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i,t}}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3}) \delta_{i_a} \right) / \sqrt{N_{i,t} \sqrt{t} \ln t} \right] \\
 &\leq 1 + 8L^5 \ln^5 T + L \ln T \sum_{t=2}^T \left[ \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i,t}}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3}) \delta_{i_a} \right) / \sqrt{N_{i,t} \sqrt{t} \ln t} \right] \\
 &\leq 1 + 8L^5 \ln^5 T + L \ln T \sum_{t=2}^T \left[ \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln C + 16 \ln t}{\delta_{i_a}} + (2 + \frac{2\pi^2}{3}) \delta_{i_a} \right) / (\sqrt{Ct} \ln t) \right] \\
 &\leq 1 + 8L^5 \ln^5 T + \frac{2L\sqrt{T} \ln T}{\sqrt{C}} \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \left( \frac{16 \ln C + 16}{\delta_{i_a}} + 9\delta_{i_a} \right), \tag{32}
 \end{aligned}$$

where we use  $\ln t \leq 2 \ln(\ln t)$  for  $t \leq 2$  in inequality (32).

Applying it to inequality (29), we can bound the expected number of selection as the following:

$$\begin{aligned}
 \mathbb{E}[N_{i,T}] &= \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + \sum_{t=n}^T \rho_{i,t} \left( \frac{4\sqrt{t} \ln t}{\Delta_i^2} \right) + \rho_{i^*,t}(\sqrt{t} \ln t) \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T \\
 &\quad + 1 + 8L^5 \ln^5 T + 2L\sqrt{T} \ln T \sum_{\substack{i_a^* \in \mathcal{O}_{i^*} \\ \mu_{i_a^*} < \mu_{i_1^*}}} \left( \frac{16}{\delta_{i_a^*}} + 9\delta_{i_a^*} \right) \\
 &\quad + 1 + 8L^5 \ln^5 T + 2\Delta_i L\sqrt{T} \ln T \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \left( \frac{16 \ln(4/\Delta_i^2) + 16}{\delta_{i_a}} + 9\delta_{i_a} \right) \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + 2 + 16L^5 \ln^5 T \\
 &\quad + 2L\sqrt{T} \ln T \sum_{\substack{i_a^* \in \mathcal{O}_{i^*} \\ \mu_{i_a^*} < \mu_{i_1^*}}} \left( \frac{16}{\delta_{i_a^*}} + 9\delta_{i_a^*} \right) + 2L\sqrt{T} \ln T \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \left( \frac{48}{\delta_{i_a}} + 9\delta_{i_a} \right),
 \end{aligned}$$

where we use the fact  $\ln(4x^2)/x \leq 2$  for any  $x \geq 0$  and  $\Delta_i \leq 1$  for the last inequality.

Hence, the expected regret from suboptimal agents can be upper bounded as the following:

$$\begin{aligned}
 \mathbb{E}[\text{Regret}_T^{\text{sub}}] &= \sum_{i \in \mathcal{N}'} \mathbb{E}[N_{i,T}] \Delta_i^m \\
 &\leq \frac{4\sqrt{T} \ln T}{\Delta_i^2} + \sqrt{T} \ln T + 2 + 16L^5 \ln^5 T \\
 &\quad + 2L\sqrt{T} \ln T \sum_{\substack{i_a^* \in \mathcal{O}_{i^*} \\ \mu_{i_a^*} < \mu_{i_1^*}}} \left( \frac{16}{\delta_{i_a^*}} + 9\delta_{i_a^*} \right) + 2L\sqrt{T} \ln T \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \left( \frac{48}{\delta_{i_a}} + 9\delta_{i_a} \right)].
 \end{aligned}$$

Combining it with inequality (26) and use the fact that any regret is at most 1, we conclude the proof.  $\square$

**Remark - Regret bound under logarithmic agent exploration.** Suppose that H-UCB is equipped with its phase 1 exploration term as  $R_i + \sqrt{2 \ln t / N_{i,t}}$ . Then, following the analogous step of our analysis, Lemma 3 should be replaced into the bounding probability with  $\sqrt{2 \ln t / N_{i,t}}$ , and this makes the bounding probability in equation (24) be  $O(\ln N_{i,t}) / O(\sqrt{N_{i,t} \ln t})$ , where one can find that this finally leads to the undesirable result of linear regret upper bound.

**Remark - Improving regret bound.** We note that there exists some tuning points in our regret analysis. For each agent  $i$ , we can improve each arm's deviation probability and the internal regret bound by replacing phase 2 UCB index to KL-UCB index [Garivier and Cappé, 2011]. Secondly, we mainly upper-bound the probability that agent  $i$  is selected given that agent  $i$  and  $i^*$  is sampled sufficiently, take expectation over it, and directly derive the regret bound by multiplying it with  $\Delta_i^m$ . In this procedure, the regret bound might be tightened if we separately upper-bound the probability that agent  $i$ 's arm  $i_a$  is selected given that agent  $i$  and  $i^*$  is sampled sufficiently, since it will lead to the separation of final regret bound by the summation of expected number of count that  $i_a$  is sampled multiplied by  $\delta_{i_a}$ .

## C.5 Proof of Theorem 6

*Proof.* To derive a problem-independent regret upper-bound, we need to remove the dependency of our regret bound from  $\delta_{i_a}$  and  $\Delta_i$  in all the denominators. To handle  $\delta_{i_a}$ , we start with the inequality (24) in Lemma 3. Let  $K = 1/c$ . For agent  $i$ , suppose that the arms are divided into the following two groups:



- (1) Group 1 consists of almost optimal arm with  $\delta_{i_a} < \sqrt{K \ln T \ln N_{i,t}/N_{i,t}}$  .  
 (2) Group 2 consists of arm with  $\delta_{i_a} \geq \sqrt{K \ln T \ln N_{i,t}/N_{i,t}}$  .

Given  $N_{i,t}$ , the total expected internal regret of agent  $i$  until  $N_{i,t}$  is the sum of the regret of each group. Since group 1 only contains near-optimal arms,

$$\sum_{i_a \in \text{Group 1}} N_{i_a,t} \delta_{i_a} < \sqrt{K \ln T \ln N_{i,t}/N_{i,t}} \sum_{i_a \in \text{Group 1}} N_{i_a,t} \leq \sqrt{K N_{i,t} \ln N_{i,t} \ln T} .$$

For the arms in group 2, the expected number of any arms in group 2 selected until round  $N_{i,t}$  is,

$$\begin{aligned} \sum_{i_a \in \text{Group 2}} \mathbb{E}[N_{i_a,t} \delta_{i_a,t}] &\leq 1 + K \ln T \left( \sum_{\substack{i_a \in \mathcal{O}_i \cap \text{Group 2} \\ \mu_{i_a} < \mu_{i_1}}} \frac{8 \ln N_{i,t}}{\delta_{i_a}} + (1 + \frac{\pi^2}{3}) \delta_{i_a} \right) \\ &\leq 8 \sqrt{K N_{i,t} \ln N_{i,t} \ln T} + 5K \ln T + 1 , \end{aligned} \quad (33)$$

where inequality (33) follows from Lemma 1 and 2. Hence, we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{a=1}^{l(i)} \delta_{i_a} N_{i_a,t} \right] &\leq 9 \sqrt{1/c N_{i,t} \ln N_{i,t} \ln T} + 5L \ln N_{i,t} + 1 \\ &\leq 9 \ln T \sqrt{K N_{i,t}} + 5K \ln T + 1 . \end{aligned}$$

Hence, inequality (25) can be replaced into the following:

$$\mathbb{P} \left[ \sum_{a=2}^{l(i)} |\mu_{i_1} - \mu_{i_a}| \cdot \frac{N_{i_a,t}}{N_{i,t}} > \sqrt{\frac{\sqrt{t} \ln t}{4N_{i,t}}} \right] \leq \frac{9 \ln T \sqrt{K N_{i,t}} + 5K \ln T + 1}{\sqrt{N_{i,t}} \sqrt{t} \ln t} .$$

On top of that, following the analogous arguments in the proof of Theorem 4 leads us to the following:

**Lemma 5.** *At round  $t$ , if any suboptimal agent  $i$  has been played at least  $\frac{4\sqrt{t} \ln t}{\Delta_i^2}$  times and optimal agent  $i^*$  has been played at least  $\sqrt{t} \ln t$ , then  $UCB_{i,t} < UCB_{i^*,t}$  with probability at least  $1 - (\rho'_{i,t}(\frac{4\sqrt{t} \ln t}{\Delta_i^2}) + \rho'_{i^*,t}(\sqrt{t} \ln t))$ , i.e.*

$$\mathbb{P}[I_{t+1} \in \mathcal{S}_i | N_{i,t} \geq \frac{4\sqrt{t} \ln t}{\Delta_i^2}, N_{i^*,t} \geq \sqrt{t} \ln t] \leq \rho'_{i,t}(\frac{4\sqrt{t} \ln t}{\Delta_i^2}) + \rho'_{i^*,t}(\sqrt{t} \ln t) , \quad (35)$$

where  $\rho'_{i,t}(x) = 1/T + 2K \ln T \exp(\frac{-\sqrt{t} \ln t}{2KN^2 \ln^2 T}) + \frac{9 \ln T \sqrt{Kx} + 5K \ln T + 1}{\sqrt{x} \sqrt{t} \ln t}$  .

We skip the detailed proof since it can similarly be derived as the proof of Lemma 4.

By integral test, the following holds:

$$\sum_{t=2}^T \frac{1}{(t \ln^2 t)^{1/4}} \leq \sum_{t=2}^T \left(\frac{1}{t}\right)^{1/4} \leq \int_{t=2}^{T+1} \frac{d(4/3t^{3/4})}{dt} dt \leq 4/3T^{3/4} . \quad (36)$$

By inequality (30) and (36), the summation of  $\rho'_{i,t}(C\sqrt{t} \ln t)$  can be bounded as the following:

$$\begin{aligned} \sum_{t=2}^T \rho'_{i,t}(C\sqrt{t} \ln t) &\leq 1 + 8K^5 \ln^5 T + \sum_{t=2}^T \left( \frac{9 \ln T \sqrt{KC\sqrt{t} \ln t} + 5K \ln T + 1}{\sqrt{Ct} \ln t} \right) \\ &\leq 1 + 8K^5 \ln^5 T + \sum_{t=2}^T \left( \frac{9 \ln T \sqrt{K}}{\sqrt{C\sqrt{t} \ln t}} + \frac{5K \ln T + 1}{\sqrt{Ct} \ln t} \right) \\ &\leq 1 + 8K^5 \ln^5 T + 12\sqrt{K/C} T^{3/4} \ln T + 10/\sqrt{CT}^{1/2} \ln T + 2\sqrt{T/C} . \end{aligned}$$

Following the analogous step to derive inequality (29) in the proof of Theorem 5, we eventually have the following inequality for any suboptimal agent  $i$ :

$$\begin{aligned}\mathbb{E}[N_{i,T}] &\leq \frac{4\sqrt{T}\ln T}{\Delta_i^2} + \frac{\sqrt{T}\ln T}{2} + \sqrt{K}T^{3/4}\ln T(12 + 6\Delta_i) + 16L^5\ln^5 T + 5T^{1/2}\ln T(2 + 1\Delta_i) + \sqrt{T}(2 + \Delta_i) \\ &\leq 18\sqrt{K}T^{3/4}\ln T + T^{1/2}\ln T\left(19 + \frac{4}{\Delta_i^2}\right) + 16L^5\ln^5 T\end{aligned}$$

Now, let's divide the suboptimal agents into two groups: For agent  $i$ , suppose that the arms are divided into the following two groups:

- (1) Group 1 consists of almost optimal agents with  $\Delta_i < (16n^2\ln^2 T/T)^{1/4}$ .
- (2) Group 2 consists of arm with  $\Delta_i \geq (16n^2\ln^2 T/T)^{1/4}$ .

Given  $N_{i,t}$ , the total expected regret is the sum of the regret in these two groups. Since group 1 only contains near-optimal agents,

$$\sum_{i \in \text{Group 1}} N_{i,T}\Delta_i \leq (16n^2\ln^2 T/T)^{1/4} \sum_{i \in \text{Group 1}} N_{i,t} \leq 2\sqrt{n}T^{3/4}\ln^{1/2} T.$$

For the agents in group 2,

$$\begin{aligned}\sum_{i \in \text{Group 2}} \mathbb{E}[N_{i,T}]\Delta_i &\leq \sum_{i \in \text{Group 2}} \left[18\sqrt{K}T^{3/4}\ln T + T^{1/2}\ln T\left(19 + \frac{4}{\Delta_i^2}\right) + 16K^5\ln^5 T\right] \\ &\leq T^{3/4}\ln T(18n\sqrt{K} + 19n + 2\sqrt{n}) + 16nK^5\ln^5 T.\end{aligned}$$

Summing the regret bound for group 1 and 2 yields,

$$\sum_{i \in N'} \mathbb{E}[N_{i,T}]\Delta_i = T^{3/4}\ln T(18n\sqrt{K} + 19n + 4\sqrt{n}) + 16nL^5\ln^5 T.$$

From inequality (34), we already have the problem-independent regret upper-bound for agent in  $N^*$  as follow,

$$\begin{aligned}\sum_{i \in N^*} \mathbb{E}\left[\sum_{a=1}^{l(i)} \delta_{i_a} N_{i_a,t}\right] &\leq \sum_{i \in N^*} \left(9\ln T\sqrt{KT} + 5K\ln T + 1\right) \\ &\leq 9n\sqrt{K}T^{1/2}\ln T + 5nK\ln T + n.\end{aligned}$$

Hence summing up the results give us the following regret bound:

$$T^{3/4}\ln T(18n\sqrt{K} + 19n + 2\sqrt{n}) + 9n\sqrt{K}T^{1/2}\ln T + 16nK^5\ln^5 T + 5nK\ln T + n = O\left(\frac{n}{\sqrt{c}}T^{\frac{3}{4}}\ln T\right),$$

and we conclude the proof.  $\square$

**Remark - honest agent.** Suppose that *honest agent*, who honestly register all the original arms without any replication, exists rather than replicators. In this case, we can get rid of the arm sampling part in RH-UCB, and correspondingly, the order of nominator in inequality (25) can be bounded by  $O(\ln N_{i,t})$ . In this case, following the analogous step provided in the analysis, we can find that eventual problem-dependent regret bound under this scenario will be  $O(\sqrt{T}\ln T)$ , and problem-independent bound will be  $O(T^{3/4}\ln^{1/4} T)$ . Note that all the analysis can also be generalized into the scenario when the strategic agent, honest agent, and replicator coexist, where the order of regret will be the same as that of the case when strategic agent and replicator exist since honest agent can be regarded as a weaker version of replicator.

## D Prior-free Robust Hierarchical UCB and its regret analysis

We provide a detailed description of Prior-free Robust Hierarchical UCB (PRH-UCB) which we briefly described in Section 4, and present corresponding regret analysis. We note that PRH-UCB still achieves replication-proofness, where we skip the proof since it can analogously be derived from the proof of Theorem 2 since it inherits the hierarchical structure of H-UCB. Note that we tune **Phase 1** exploration parameter to  $\sqrt{t \ln^3 t}$  to optimize the order of poly-logarithmic term in regret bound. For the regret analysis, we only provide a proof sketch since the detailed steps are analogous to the proof of Theorem 5 and 6.

**Theorem 7.** *PRH-UCB has  $O(\frac{n}{c^2} T^{\frac{1}{2}} \ln^{\frac{3}{2}} T)$  problem-dependent and  $O(\frac{n}{\sqrt{c}} T^{\frac{3}{4}} \ln^{\frac{3}{2}} T)$  problem-independent regret upper-bound for any  $T$  such that  $T \ln^3 T \geq e^{2/c}$ .*

---

### Algorithm 3: Prior-free Robust Hierarchical UCB (PRH-UCB)

---

```

1 Initialize  $R(i), N(i), r(a)$  and  $n(a)$  at zero and  $\mathcal{B}_i$  at an empty set for all  $i \in \mathcal{N}$  and  $a \in \cup_{i \in \mathcal{N}} \mathcal{S}_i$ ; for
   $t = 1, 2, \dots$  do
  | // Phase 1 - agent selection
  | if Unexplored agent exists then
  | | Pick an unexplored  $\hat{i} \in \{i \in \mathcal{N} : N(i) = 0\}$ ;
  | else
  | | Pick  $\hat{i} = \operatorname{argmax}_{i \in \mathcal{N}} \left\{ R(i) + \sqrt{\frac{\sqrt{t \ln^3 t}}{N(i)}} \right\}$ ;
  | // Phase 2 - arm selection
  | if  $|\mathcal{B}_i| < \min(|\mathcal{S}_i|, \ln^2 t)$  then
  | | Pick an unexplored arm  $\hat{a} \in \{a \in \mathcal{S}_i \setminus \mathcal{B}_i : n(a) = 0\}$ ;
  | |  $\mathcal{B}_i \leftarrow \mathcal{B}_i \cup \{\hat{a}\}$ ;
  | else
  | | Pick  $\hat{a} = \operatorname{argmax}_{a \in \mathcal{B}_i} \left\{ r(a) + \sqrt{\frac{2 \ln N(i)}{n(a)}} \right\}$ ;
  | Play  $A_t = \hat{a}$  and receive reward  $R_t$ ;
  | Update statistics:  $r(\hat{a}) \leftarrow \frac{r(\hat{a})n(\hat{a}) + R_t}{n(\hat{a}) + 1}$ ;  $R(\hat{i}) \leftarrow \frac{R(\hat{i})N(\hat{i}) + R_t}{N(\hat{i}) + 1}$ ;  $n(\hat{a}) \leftarrow n(\hat{a}) + 1$ ;  $N(\hat{i}) \leftarrow N(\hat{i}) + 1$ ;
  
```

---

*Proof sketch.* The proof structure is exactly the same with the proof of Theorem 5 and Theorem 6. We briefly explain how the order of magnitude will be changed when the sampling amount becomes  $\ln^2 T$ . Given round  $t$ , we redefine the random variable  $Y_i$  to be the event that  $\mathcal{B}_i$  contains  $i$ 's optimal arm or any replica of it until round  $t$ . Then we can derive the following lemma:

**Lemma 6.** *At round  $t$ , if  $N_{i,t} \geq e^{1/c}$ , then  $\mathbb{P}[Y_i^c] \leq 1/N_{i,t}$ .*

*Proof of Lemma 6.* At round  $t$ , agent  $i$  is selected  $N_{i,t}$  times and  $\min(|\mathcal{S}_i|, \ln^2 N_{i,t})$  arms will be sampled until then. If  $|\mathcal{S}_i| \leq \ln^2 N_{i,t}$ , then the optimal arm will be contained in  $\mathcal{B}_i$ . Otherwise, the following inequalities conclude the proof.

$$\begin{aligned}
 \mathbb{P}[Y_i^c] &\leq (1-c) \frac{|\mathcal{S}_i|(1-c)-1}{|\mathcal{S}_i|-1} \frac{|\mathcal{S}_i|(1-c)-2}{|\mathcal{S}_i|-2} \cdots \frac{|\mathcal{S}_i|(1-c)-\ln^2 N_{i,t}}{|\mathcal{S}_i|-\ln^2 N_{i,t}} \\
 &\leq (1-c)^{\ln^2 N_{i,t}} \\
 &\leq (1-c)^{1/c \ln N_{i,t}} \\
 &\leq 1/N_{i,t},
 \end{aligned}$$

where the last inequality holds from  $(1 - 1/x)^x \leq 1/e$  for any  $x \geq 0$ . □

Then, Lemma 2 should be replaced as the following:

**Lemma 7.** *At round  $t$ , conditional expected internal regret given that  $Y_i$  is occurred can be bounded as the following:*

$$\sum_{i_a \in \mathcal{B}_i} \mathbb{E}[\delta_{i_a} N_{i_a, t} | Y_i] \leq \ln^2 N_{i, t} \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{8 \ln N_{i, t}}{\delta_{i_a}} + \left(1 + \frac{\pi^2}{3}\right) \delta_{i_a} \right).$$

*Proof of Lemma 7.* Given  $t$  and  $N_{i, t}$ ,  $|\mathcal{B}_i| = \min(|\mathcal{S}_i|, \ln^2 N_{i, t})$ . If  $|\mathcal{S}_i| \leq \ln^2 N_{i, t}$ , then the optimal arm will be contained in  $\mathcal{B}_i$ , and the result directly holds. Otherwise, we have

$$\mathbb{E}[\delta_{i_a} N_{i_a, t} | Y_i] \leq (1 - \mathbb{P}[Y_i]) \ln^2 N_{i, t} \left( \sum_{\substack{a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \left( \frac{8 \ln N_{i, t}}{\delta_{i_a}} + \left(1 + \frac{2\pi^2}{3}\right) \delta_{i_a} \right) \right),$$

and it concludes the proof.  $\square$

Now we present two lemmas which are analogous to Lemma 3 and 4.

**Lemma 8.** *Given any agent  $i \in \mathcal{N}$ , for each arm  $j$  in  $j \in \mathcal{S}_i$  at time  $t$ , we have*

$$\mathbb{P} \left[ |R_{i, t} - \mu_{i_1}| \geq \sqrt{\frac{\sqrt{t \ln^3 t}}{N_{i, t}}} \right] \leq \rho''(N_{i, t}),$$

where

$$\rho''(N_{i, t}) = 1/N_{i, t} + \frac{1}{c} \ln T e^{-\sqrt{t \ln^3 t}/(2L^2 \ln^2 T)} + \frac{1}{c} \ln^2 N_{i, t} \left( \sum_{\substack{i_a \in \mathcal{O}_i \\ \mu_{i_a} < \mu_{i_1}}} \frac{16 \ln N_{i, t}}{\delta_{i_a}} + \left(2 + \frac{2\pi^2}{3}\right) \delta_{i_a} \right) / \sqrt{N_{i, t} \sqrt{t \ln^3 t}}.$$

**Lemma 9.** *At round  $t$  such that  $t \ln^3 t \geq e^{2/c}$ , if any suboptimal replicator  $i$  has been played at least  $\frac{4\sqrt{t \ln^3 t}}{\Delta_i^2}$  times and optimal replicator  $i^*$  has been played at least  $\sqrt{t \ln^3 t}$ , then  $UCB_{i, t} < UCB_{i^*, t}$  with probability at least  $1 - (\rho''_{i, t}(\frac{4\sqrt{t \ln^3 t}}{\Delta_i^2}) + \rho''_{i^*, t}(\sqrt{t \ln^3 t}))$ , i.e.*

$$\mathbb{P} \left[ I_{t+1} \in \mathcal{S}_i | N_{i, t} \geq \frac{4\sqrt{t \ln^3 t}}{\Delta_i^2}, N_{i^*, t} \geq \sqrt{t \ln^3 t} \right] \leq \rho''_{i, t} \left( \frac{4\sqrt{t \ln^3 t}}{\Delta_i^2} \right) + \rho''_{i^*, t}(\sqrt{t \ln^3 t}). \quad (37)$$

Then, followed by some calculations, one can find that  $\rho''(\sqrt{t \ln^3 t}) = O(\sqrt{t \ln^3 t})$  for any  $t$  such that  $t \ln^3 t \geq e^{2/c}$ . These enables us to obtain  $O(\frac{n}{c^2} \sqrt{T \ln^3 T})$  problem-dependent regret upper bound and  $O(\frac{n}{\sqrt{c}} T^{3/4} \ln^{3/2} T)$  problem-independent regret upper bound. Note that deriving problem-independent regret bound requires a similar technique as the ones used in the proof of Theorem 6.  $\square$

## E Sampling UCB

We provide a pseudo-code for S-UCB described in Section 5, which is UCB1 with the subsampling module of RH-UCB. Our construction of S-UCB is to perform an ablation study on the impact of subsampling and hierarchy in RH-UCB, respectively. We recall that our experiment reveals that S-UCB can outperform UCB1 in cases with infinite replications thanks to the addition of subsampling module, c.f., Figure 2(b) and Figure 2(c).

---

### Algorithm 4: Sampling UCB (S-UCB)

---

```

1 Parameter:  $L > 0$ 
2 Initialize  $R(i), N(i), r(a)$  and  $n(a)$  at zero for all  $i \in \mathcal{N}$  and  $a \in \cup_{i \in \mathcal{N}} \mathcal{S}_i$ ;
3 Subsample  $\mathcal{B}_i$  of size  $\min\{|\mathcal{S}_i|, L \ln T\}$  from  $\mathcal{S}_i$  and  $\mathcal{B} = \cup_i \mathcal{B}_i$ ;
4 for  $t = 1, 2, \dots$  do
5   if Unexplored arm exists in  $\mathcal{B}$  then
6     | Pick an unexplored  $\hat{a} \in \{a \in \mathcal{B} : n(a) = 0\}$ ;
7   else
8     | Pick  $\hat{a} = \operatorname{argmax}_{a \in \mathcal{B}} \left\{ r(a) + \sqrt{\frac{2 \ln t}{n(a)}} \right\}$ ;
9   end
10  Play  $A_t = \hat{a}$  and receive reward  $R_t$ ;
11  Update statistics:  $r(\hat{a}) \leftarrow \frac{r(\hat{a})n(\hat{a}) + R_t}{n(\hat{a}) + 1}$ ;  $n(\hat{a}) \leftarrow n(\hat{a}) + 1$ ;
12 end

```

---