Competing for Shareable Arms in Multi-Player Multi-Armed Bandits

Renzhe Xu [*1] **Haotian Wang** [*1] **Xingxuan Zhang** [1] **Bo Li** [2] **Peng Cui** [1]

Abstract

Competitions for shareable and limited resources have long been studied with strategic agents. In reality, agents often have to learn and maximize the rewards of the resources at the same time. To design an individualized competing policy, we model the competition between agents in a novel multi-player multi-armed bandit (MPMAB) setting where players are selfish and aim to maximize their own rewards. In addition, when several players pull the same arm, we assume that these players averagely share the arms' rewards by expectation. Under this setting, we first analyze the Nash equilibrium when arms' rewards are known. Subsequently, we propose a novel **S**elfish **M**PMAB with **A**veraging **A**llocation (SMAA) approach based on the equilibrium. We theoretically demonstrate that SMAA could achieve a good regret guarantee for each player when all players follow the algorithm. Additionally, we establish that no single selfish player can significantly increase their rewards through deviation, nor can they detrimentally affect other players' rewards without incurring substantial losses for themselves. We finally validate the effectiveness of the method in extensive synthetic experiments.

1. Introduction

Agents compete with each other to maximize their own rewards in various applications with shareable and limited resources, ranging from cognitive radio networks (Mitola & Maguire, 1999) to algorithm-curated platforms (Hron et al., 2022). For example, secondary users compete for channels to transmit data in cognitive radio networks (Mitola & Maguire, 1999; Haykin, 2005; Meshkati et al., 2007). Regarding algorithm-curated platforms (such as YouTube and TikTok), content providers create content with different topics to compete for exposure (Ben-Porat et al., 2020; Hron et al., 2022; Jagadeesan et al., 2022; Yao et al., 2023; Zhu et al., 2023). In these applications, resources are often shareable, and each resource has a limited total reward, such as the quality of each channel in cognitive radio networks and the demand for each content topic in algorithm-curated platforms. As a result, when more agents choose to acquire the same resource, the reward earned by each of them will decrease. Meanwhile, agents usually do not have prior knowledge about the total reward of each resource in reality (Jouini et al., 2009; 2010; Li et al., 2010). Therefore, these agents need to design policies to maximize their rewards faced with unknown resource rewards and competition from other agents.

It is a common practice to adopt the multi-player multi-armed bandit (MPMAB) framework (Anantharam et al., 1987; Liu & Zhao, 2010; Anandkumar et al., 2011; Boursier & Perchet, 2022) to model the behaviors of multiple agents with unknown rewards of different resources. In this framework, players pull arms simultaneously at various rounds. Since several players may choose the same arm, different collision models (Bande & Veeravalli, 2019; Liu et al., 2020; Shi & Shen, 2021; Boyarski et al., 2021; Wang et al., 2022a;b) are proposed to allocate the arm's reward to collided players. However, these works mainly focus on the cooperative player setting where players are non-strategic and cooperate to maximize the total reward of all players. This does not agree with the real scenarios since players are always selfish and target to maximize their own reward. Several works (Boursier & Perchet, 2020; Liu et al., 2020; 2021; Jagadeesan et al., 2021) consider the selfish behaviors of players while they assume that the resource is not shareable, as at most one player can get the reward when a collision occurs.

In this paper, we propose a novel MPMAB setting to characterize the strategic behaviors of agents in applications with limited and shareable resources. We assume that the players are selfish, and they could strategically alter their policies to achieve better rewards. In addition, when several players pull the same arm, we suppose that these players

[*]Equal contribution [1]Department of Computer Science and Technology, Tsinghua University, Beijing, China [2]School of Economics and Management, Tsinghua University, Beijing, China. Emails: xrz199721@gmail.com, accwht@hotmail.com, xingxuanzhang@hotmail.com, libo@sem.tsinghua.edu.cn, cuip@tsinghua.edu.cn. Correspondence to: Peng Cui <cuip@tsinghua.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

averagely share the arms' rewards by expectation. We further assume that each player could observe his own reward and the reward of the chosen arm, similar to the statistic sensing setting in traditional MPMAB literature (Boursier & Perchet, 2022). Our target is to design an algorithm for each player that maximizes his own reward and is robust to the strategic behaviors of any other player. Specifically, the algorithm should achieve a good regret guarantee when all players follow the algorithm. Furthermore, no single selfish player can substantially increase their rewards through deviation, nor can they negatively impact other players' rewards without incurring significant personal losses.

To achieve this target, we first analyze the Nash equilibrium of players' behaviors when the expected reward of each arm is known. Based on the equilibrium, we propose a novel **S**elfish **M**PMAB with **A**veraging **A**llocation (SMAA) approach for each player that can explore the rewards of arms, maximize his own reward, and converge to equilibrium at the same time. We theoretically analyze our algorithm and demonstrate that after T rounds, (1) when all players follow SMAA, the regret for each player is $O(\log T)$. We further show that it matches the lower bound when players only use the information of chosen arms' rewards. (2) When all players follow SMAA, the algorithm could converge to equilibrium, and the number of non-equilibrium rounds is $O(\log T)$. (3) A selfish player that deviates from the algorithm will bring at most $O(\log T)$ increase in his own reward. (4) If a selfish player wants to bring loss u to another player's reward, he will also suffer from a loss of at least $\beta u - O(\log T)$, where β is a constant. We finally validate the effectiveness of our method through extensive synthetic experiments.

To conclude, our contributions are listed as follows.

- We propose a novel MPMAB setting with an averaging allocation model to characterize the selfish behaviors of players in applications with limited and shareable resources.
- We analyze the Nash equilibrium of the problem and further propose a novel Selfish MPMAB with Averaging Allocation (SMAA) approach for players under the proposed setting.
- We theoretically demonstrate that SMAA could achieve a good regret guarantee for each player when all players follow the algorithm and it is robust to a single player's strategic deviation.

2. Preliminaries

2.1. Basic Setting

Suppose there are K arms with indices $[K] = \{1, 2, \ldots, K\}$, N players (he) and T rounds. At round $t \in [T]$, the reward of arm k is given by $X_k(t)$ drawn *i.i.d.* according to a distribution $\xi_k \in \Xi$ with expectation $\mu_k > 0$ supported on $[0, 1]$. Here Ξ represents a set of possible distributions supported on $[0, 1]$ with positive expectations.

The problem is decentralized. At each round $t \in [T]$, N players choose arms simultaneously based only on their historical observations. We use $\pi_j(t)$ to denote the arm chosen by player j and $M_k(t) = \sum_{j=1}^{N} \mathbb{I}[\pi_j(t) = k]$ to represent the number of players that pull arm k at round t.

We assume all players are homogeneous, leading to an averaging allocation mechanism when collisions occur. Formally, at round t, each player j is endowed with a weight $w_j(t)$ sampled *i.i.d.* from a distribution Γ supported on $[0, 1]$. Then each arm's reward is allocated proportionally to the players' weights. Let random variable $R_j(t)$ denote the reward earned by player j at round t, and it is given by the following equation:

$$R_j(t) = X_{\pi_j(t)}(t) \cdot \frac{w_j(t)}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'}(t) = \pi_j(t)] w_{j'}(t)}. \quad (1)$$

As a result, since all players' weights are sampled from the same distribution Γ, we have that for all $k \in [K]$ such that $M_k(t) > 0$ and $j \in [N]$ such that $\pi_j(t) = k$,

$$\mathbb{E}\left[R_j(t) | X_k(t), M_k(t)\right] = X_k(t)/M_k(t). \quad (2)$$

This indicates that players averagely share the arms' rewards by expectation.

We consider the setting similar to the statistic sensing setting in standard multi-player multi-armed bandit problems (Boursier & Perchet, 2022). Formally, at each round, player j could observe his own reward $R_j(t)$ and the total reward of the selected arm $\pi_j(t)$ (*i.e.*, $X_{\pi_j(t)}(t)$).

Example 2.1. We demonstrate one example of our proposed problem setting. In an algorithm-curated platform (Hron et al., 2022), such as YouTube and TikTok, content providers create content with different topics to compete for exposure. Here content providers are players and different content topics are arms. The reward of each arm corresponds to the total exposure to the topic. As the demand for each topic is usually stable (Ben-Porat et al., 2020; Hron et al., 2022), the expected amount of exposure for each topic is unchanged. Therefore, assuming players are homogeneous (Hron et al., 2022; Jagadeesan et al., 2022), when several content providers select the same topic, the exposure is shared averagely among these providers by expectation.

Other mathematical notations We use $\lfloor x \rfloor$ to denote the largest integer that is not greater than x and $\lceil x \rceil$ to denote the smallest integer that is not smaller than x for any real number x. For any $(p, q) \in [0, 1]^2$, let $\mathrm{kl}(p, q)$ denote the Bernoulli Kullback-Leibler divergence defined as $\mathrm{kl}(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$

(Garivier & Cappé, 2011). In addition, with convention, we have that $0 \log 0 = 0 \log 0/0 = 0$ and $x \log x/0 = \infty$ if $x > 0$.

Background on Nash equilibrium We adopt the definition of ϵ-Nash equilibrium from the game theory (Nisan et al., 2007). For a general game problem, let $\mathcal{S}$ be the set of agents' strategies, and each agent j follows a strategy $s_j \in \mathcal{S}$. We use $\boldsymbol{s} = \{s_j\}_{j=1}^N$ to denote the strategy profile of all players and $(s', \boldsymbol{s}_{-j})$ to denote the profile where a single player j deviates from the original strategy s_j to a new strategy $s' \in \mathcal{S}$. In addition, we use $\mathcal{U}_j(\boldsymbol{s})$ to denote the utility of player j when the strategy profile is $\boldsymbol{s}$. With these notations, ϵ-Nash equilibrium is defined as follows.

Definition 2.1 (ϵ-Nash equilibrium)**.** For a game specified by $(\mathcal{S}, \{\mathcal{U}_j\}_{j=1}^N)$, a strategy profile $\boldsymbol{s} \in \mathcal{S}^N$ is an ϵ-Nash equilibrium if for all $s' \in \mathcal{S}$ and $j \in [N]$,

$$\mathbb{E}\left[\mathcal{U}_j(s', \boldsymbol{s}_{-j})\right] \leq \mathbb{E}[\mathcal{U}_j(\boldsymbol{s})] + \epsilon. \tag{3}$$

Moreover, if a strategy profile $\boldsymbol{s}$ is a 0-Nash equilibrium, we say that the strategy profile $\boldsymbol{s}$ is a Nash equilibrium.

2.2. Evaluation metrics

In this paper, we target an algorithm for each player that maximizes his own reward and is robust to the strategic behaviors of any other player. We first introduce two metrics to evaluate a policy's performance when all players follow the algorithm.

Players' regret Since we focus on the selfish player setting, we evaluate the regret of each player compared with his best choice in each round. At round t, if player j pulls arm $k \in [K] \backslash \{\pi_j(t)\}$, the expected reward is $\mu_k/(M_k(t)+1)$. If player j pulls the originally chosen arm $\pi_j(t)$, his expected reward is $\mu_{\pi_j(t)}/M_{\pi_j(t)}(t)$. Combining these two cases, we can get the following equation for player j's regret.

$$\mathrm{Reg}_j(T) = \sum_{t=1}^{T} \left(\max_{k \in [K]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t) \right). \tag{4}$$

Notably, each agent's regret depends on others agents' policies since $M_k(t)$ is the number of players on arm k.

Number of non-equilibrium rounds When players know the arms' expected rewards, their chosen arms will follow the Nash equilibrium demonstrated in Section 3.1. As a result, we use the number of non-equilibrium rounds to evaluate whether an algorithm could effectively converge. Formally, the metric can be calculated by the following equation.

$$\mathrm{NonEqu}(T) = \sum_{t=1}^{T} \mathbb{I}\left[\exists k \in [K], M_k(t) \neq m_k^*\right], \tag{5}$$

where m_k^* is provided in Theorem 3.1 and it represents the number of players that pull arm k in the equilibrium.

We further provide two perspectives to analyze the policy's stability when there exists a single strategic player. We use $\mathcal{S}_{\text{policy}}$ to denote the set of all policies for players. For any $\boldsymbol{s} \in \mathcal{S}_{\text{policy}}^N$, the utility of player j is the total reward after T rounds, *i.e.*, $\mathcal{U}_j^{\text{policy}}(\boldsymbol{s}) = \mathrm{Rew}_j(T; \boldsymbol{s}) = \mathbb{E}\left[\sum_{t=1}^T R_j(t; \boldsymbol{s})\right]$. Here we slightly abuse the notation and use $R_j(t; \boldsymbol{s})$ to represent the reward earned by player j at round t when the policy profile is $\boldsymbol{s}$.

ϵ-Nash equilibrium *w.r.t.* policies We expect a policy profile $\boldsymbol{s}$ is a ϵ-Nash equilibrium *w.r.t.* the game specified by $(\mathcal{S}_{\text{policy}}, \{\mathcal{U}_j^{\text{policy}}\}_{j=1}^N)$ with a small ϵ such that any single selfish player can not significantly improve his own reward by deviation.

Remark 2.1. We note that the equilibrium mentioned here is not the equilibrium introduced when calculating the number of non-equilibrium rounds. Here the strategy set $\mathcal{S}_{\text{policy}}$ represents all policies and the utility function is the total reward after T rounds. However, in the equilibrium mentioned in Equation (5) (Details are provided in Section 3.1), the strategy set is the set of all arms and the utility function is the expected reward at a single round.

(β, ϵ)-stable Following (Boursier & Perchet, 2020), we further analyze how a single selfish player could affect other players' reward by the following definition.

Definition 2.2 ((β, ϵ)-stable (Boursier & Perchet, 2020))**.** A policy profile $\boldsymbol{s} \in \mathcal{S}_{\text{policy}}^N$ is (β, ϵ)-stable if for any $s' \in \mathcal{S}_{\text{policy}}$, $u \in \mathbb{R}^+$, and $i, j \in [N]$,

$$\begin{aligned} &\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j})] \leq \mathbb{E}[\mathrm{Rew}_i(T; \boldsymbol{s})] - u \\ \Longrightarrow &\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j})] \leq \mathbb{E}[\mathrm{Rew}_j(T; \boldsymbol{s})] + \epsilon - \beta u. \end{aligned} \tag{6}$$

Intuitively, if player j wants to incur a considerable loss u to player i, then he will also suffer from a comparable loss of at least $\beta u - \epsilon$.

3. Proposed Methods

We first analyze the Nash equilibrium when the arms' expected rewards are known in Section 3.1. Afterward, based on the equilibrium, we provide the novel Selfish MPMAB with Averaging Allocation (SMAA) approach for players in Section 3.2. We theoretically analyze the property of the algorithm in Section 3.3. In these subsections, we assume that each player knows the number of players and is endowed with a different rank (*i.e.*, a unique numbering from 1 to N for each agent). Finally, in Section 3.4, we relax the assumption and demonstrate that combining the classic Musical Chairs approach (Rosenski et al., 2016) with SMAA could achieve similar theoretical results.

Basic assumption Previous works on the zero collision reward setting usually assume that the expected rewards of arms are different (Boursier & Perchet, 2020; Wang et al., 2020). Extending to the shareable arm case, we slightly strengthen the assumption as shown in Assumption 3.1.

Assumption 3.1. For any $k, k' \in [K]$ and $n, n' \in [N]$, we have $\mu_k/n \neq \mu_{k'}/n'$.

Remark 3.1. We further demonstrate the rationality of the assumption by proving that it holds with probability 1 when the expected rewards of different arms μ_k are sampled randomly from an absolutely continuous probability distribution with bounded probability densities (such as uniform and beta distributions). Details can be found in Section A.1.

3.1. Nash Equilibrium When Rewards Are Known

We first analyze the Nash equilibrium at each round when players know the expected reward of each arm. Since players are selfish and aim to target their rewards, their strategies at each round will eventually converge to the Nash equilibrium.

At each round, the strategy set for each player is the set of all arms, *i.e.*, $\mathcal{S}_{\text{single}} = [K]$. For a strategy profile $\boldsymbol{s} \in \mathcal{S}_{\text{single}}^N$, we use $\boldsymbol{m}(\boldsymbol{s}) = \{m_k(\boldsymbol{s})\}_{k=1}^K$ to denote the number of players that choose arm k. Then the utility of player j is the expected reward at this round, which is given by $\mathcal{U}_j^{\text{single}}(\boldsymbol{s}) = \mu_{s_j}/m_{s_j}(\boldsymbol{s})$ according to the averaging allocation mechanism.

Our problem setting at each round is a specific instance of the singleton congestion game (SCG) (Ieong et al., 2005; Basat et al., 2017). Indeed, Ieong et al. (2005) has developed polynomial-time algorithms to discover the Nash equilibrium that corresponds to the maximal social welfare in SCGs. However, the algorithm is designed for general SCGs, failing to characterizing the detailed properties of our problem. As result, it pose significant challenges to adopt them in online settings where each arms' rewards remain unknown. To address this, we develop the following property for the Nash equilibria in our problem setting, which can be seamlessly integrated within online settings.

Theorem 3.1. *Under Assumption 3.1, the Nash equilibrium for the game specified by* $(\mathcal{S}_{\text{single}}, \{\mathcal{U}_j^{\text{single}}\}_{j=1}^N)$ *exists. Moreover, for any strategy profile* $\boldsymbol{s}$*, it is a Nash equilibrium if and only if* $\boldsymbol{m}(\boldsymbol{s}) = \boldsymbol{m}^*$ *where* $\boldsymbol{m}^*$ *is given by*

$$\boldsymbol{m}^* = \left\{ m_k^* = \left\lfloor \frac{\mu_k}{z^*} \right\rfloor \right\}_{k=1}^K, \tag{7}$$

$$\textit{where} \quad z^* = \sup\left\{ z > 0 : h(z) \triangleq \sum_{k=1}^K \left\lfloor \frac{\mu_k}{z} \right\rfloor \geq N \right\}. \tag{8}$$

We further use $\mathcal{M}^* = \{k \in [K] : m_k^* > 0\}$ to represent the arms chosen in the equilibrium.

Remark 3.2. As shown in Section A.2, the equilibria described in Theorem 3.1 are actually strong Nash equilibria, hence robust to multiple players' deviations.

Remark 3.3. Although we only consider pure strategies in Theorem 3.1 (*i.e.*, pure Nash equilibrium (PNE)), our evaluation of the randomized strategies in Section A.3 reveals that the total welfare (measured as the total rewards of all players) from any symmetric Mixed Nash equilibrium (MNE) (*i.e.*, all players follow the same randomized strategy) does not exceed that from equilibria in Theorem 3.1. We analyze symmetric MNE due to the computational difficulties inherent in general mixed Nash equilibria (Nisan et al., 2007), aligning with prior research in algorithm-curated platforms (Jagadeesan et al., 2022).

We present an example to better explain the intuition behind Theorem 3.1.

Example 3.1. Suppose there are $K = 3$ arms with expected rewards $\{\mu_1, \mu_2, \mu_3\} = \{1, 0.4, 0.2\}$ and $N = 3$ players. Then in a Nash equilibrium, two players pull arm 1 with reward 1 and one player pulls arm 2 with reward 0.4.

Theorem 3.1 and Example 3.1 demonstrate that when players reach an equilibrium, the number of players that choose each arm is approximately proportional to the corresponding arms' expected rewards. However, this may harm the total welfare (measured as the total rewards of all players). In Example 3.1, the total welfare of all players in the equilibrium is 1.4. By contrast, if a player deviates from arm 1 to arm 3, the total welfare will increase to 1.6. This phenomenon is well known as the existence of the price of anarchy (PoA) (Roughgarden, 2005; Nisan et al., 2007) in game theory, which indicates the mismatch between agents' utilities and the total welfare.

Analysis of the PoA We further analyze the bound on the value of PoA. Formally, PoA is measured as the ratio between the maximal total welfare achieved by any strategy profile and the total welfare in the equilibrium.

Proposition 3.2. *PoA is bounded between* 1 *and* $(N + \min\{K, N\} - 1)/N$*. Furthermore, PoA can take value at* 1 *and any close to* $(N + \min\{K, N\} - 1)/N$*.*

Remark 3.4. Basat et al. (2017) analyze the PoA of the same setting when $N = 2$ and demonstrate that PoA is at most 1.5, which is consistent with the proposition when $K \geq 2$. Furthermore, Proposition 3.2 extends their results to more general scenarios where N may exceed 2. Furthermore, we would like to highlight that Theorem 3.1 proves that the Nash equilibrium at each round in our proposed setting is unique with respect to the number of agents on each arm. As a result, all of the Nash equilibria in Theorem 3.1 have the same PoA and total social welfare in our setting. Details and further analysis can be found in Section A.4.

3.2. Selfish MPMAB with Averaging Allocation (SMAA)

In this subsection, we present our novel SMAA approach based on the property of the Nash equilibrium demonstrated in Theorem 3.1. Our SMAA approach is inspired by the alternate exploration method (Wang et al., 2020) while we need to estimate the Nash equilibrium at each round, which becomes more complicated.

3.2.1. Empirical Statistics

Consider a player with rank j. Let $\tau_{j,k}(t) = \sum_{s=1}^{t} \mathbb{I}[\pi_j(t) = k]$ be the number of rounds that player j chooses arm k. The estimation $\hat{\mu}_{j,k}(t)$ of player j on the reward of arm k at round t is given by

$$\hat{\mu}_{j,k}(t) = \frac{1}{\tau_{j,k}(t)} \sum_{s=1}^{t} \mathbb{I}[\pi_j(t) = k] X_k(t). \tag{9}$$

Exploration is based on the KL-UCB indices (Garivier & Cappé, 2011). The index of arm k for player j at round t is given by

$$\hat{b}_{j,k}(t) = \sup \left\{ q \geq \hat{\mu}_{j,k}(t) : \tau_{j,k}(t)\, \mathrm{kl}(\hat{\mu}_{j,k}(t), q) \leq f(t) \right\}, \tag{10}$$

where $f(t) = \log t + 4 \log(\log t)$.

At round t, the estimated equilibrium $\hat{\boldsymbol{m}}_j(t) = \{\hat{m}_{j,k}(t)\}_{k=1}^{K}$ is calculated by Equations (7) and (8) according to the estimated rewards $\{\hat{\mu}_{j,k}(t)\}_{k=1}^{K}$. In addition, let $\hat{\mathcal{M}}_j(t) = \{k \in [K] : \hat{m}_{j,k}(t) > 0\}$ be the arms in the estimated equilibrium $\hat{\boldsymbol{m}}_j(t)$. For any arm $k \in \hat{\mathcal{M}}_j(t)$, let $\hat{r}_{j,k}(t) = \hat{\mu}_{j,k}(t)/\hat{m}_{j,k}(t)$ denote the expected average reward to choose arm k in the estimated equilibrium $\hat{\boldsymbol{m}}_j(t)$ with estimated rewards $\hat{\mu}_{j,k}(t)$.

3.2.2. Algorithm Details

The pseudo-code is shown in Algorithm 1 and the algorithm consists of two phases.

Initialization phase Each player first conducts an initialization phase with $K' = N \lceil K/N \rceil$ rounds. Since $K' \geq K$ by construction, each player can pull each arm at least one time. We set the rounds for initialization as K' instead of K because $K' \equiv 0 (\mathrm{mod}\, N)$, which matches the second exploration-exploitation phase that is conducted in blocks with N rounds each.

Exploration-exploitation phase Afterward, each player enters the exploration-exploitation phase. The whole procedure is conducted in blocks with N rounds each. Generally speaking, in each block, each player sequentially pulls the arms in the estimated Nash equilibrium based on his historical observations unless the player pulls the arm with the smallest average reward for the last time. In this case, he will explore other arms with probability $1/2$.

Algorithm 1 Selfish MPMAB with Averaging Allocation (SMAA)

1: **Input:** Player rank j
2: Let $K' \leftarrow N \cdot \lceil K/N \rceil$
3: Initialization phase $1 \leq t \leq K'$:
4: Pull each arm at least one time.
5: Exploration-exploitation phase $t > K'$:
6: **for** $t \leftarrow K' + 1$ to T **do**
7: Calculate $\tilde{\tau}_{j,k}(t)$, $\tilde{\mu}_{j,k}(t)$, $\tilde{b}_{j,k}(t)$, $\tilde{\mathcal{M}}_j(t)$, $\tilde{\boldsymbol{m}}_j(t)$, $\tilde{r}_{j,k}(t)$ according to Equation (11)
8: Calculate the arms chosen in this block $\tilde{l}_j(t)$ according to Equation (12)
9: Calculate the index $i \leftarrow (t + j) \bmod N + 1$
10: Let k be the i-th element in $\tilde{l}_j(t)$
11: **if** $i = N$ **then**
12: Calculate $\mathcal{H}_j$ according to Equation (13)
13: **if** $\mathcal{H}_j(t) = \emptyset$ **then**
14: $\pi_j(t) \leftarrow k$
15: **else**
16: With probability $\frac{1}{2}$, $\pi_j(t) \leftarrow k$
17: With probability $\frac{1}{2}$, $\pi_j(t) \leftarrow k'$ chosen uniformly at random in $\mathcal{H}_j(t)$
18: **end if**
19: **else**
20: $\pi_j(t) \leftarrow k$
21: **end if**
22: Pull arm $\pi_j(t)$
23: **end for**

Formally, the behaviors of player j in a block are based on the estimations up to the end of the last block. As a result, the policy at round t is determined by the following estimations (with the tilde notation).

$$\begin{aligned}
&\tilde{\tau}_{j,k}(t) = \tau_{j,k}\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right), \tilde{\mu}_{j,k}(t) = \hat{\mu}_{j,k}\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right), \\
&\tilde{b}_{j,k}(t) = \hat{b}_{j,k}\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right), \tilde{\mathcal{M}}_j(t) = \hat{\mathcal{M}}_j\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right), \\
&\tilde{\boldsymbol{m}}_j(t) = \hat{\boldsymbol{m}}_j\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right), \tilde{r}_{j,k}(t) = \tilde{r}_{j,k}\left(\left\lfloor \frac{t-1}{N} \right\rfloor N\right).
\end{aligned} \tag{11}$$

Here the value of $\lfloor (t-1)/N \rfloor$ is the last round in the previous block.

For each round t, player j first calculates the list of arms to choose in the corresponding block as shown in Line 7-8 in Algorithm 1. In particular, player j sorts the arms in the estimated equilibrium $\tilde{\mathcal{M}}_j(t)$ according to the average reward $\tilde{r}_{j,k}(t)$ by descending order and get the indices of the arms $k_1, k_2, \ldots, k_{|\tilde{\mathcal{M}}_j(t)|}$. Then he sequentially aligns these arms and each arm k_i is further repeated by $\tilde{m}_{j,k_i}$ times. Formally, the arms chosen in the corresponding block are

given by

$$\tilde{l}_j(t) = \left(\underbrace{k_1, \dots, k_1}_{\tilde{m}_{j,k_1}(t) \text{ times}}, \dots, \underbrace{k_a, \cdots, k_a}_{\tilde{m}_{j,k_a}(t) \text{ times}} \right), \tag{12}$$

where $a = |\tilde{\mathcal{M}}_j(t)|$. We demonstrate how $\tilde{l}_j(t)$ is calculated through the numerical showcase provided in Example 3.1.

Example 3.2. Suppose the estimated expected reward of player j at a round t is $\{\tilde{\mu}_{j,k}(t)\}_{k=1}^K = \{1, 0.4, 0.2\}$. Then the estimated number of players on each arm in the equilibrium $\{\tilde{m}_{j,k}(t)\}_{k=1}^K = \{2, 1, 0\}$, and we have $\tilde{\mathcal{M}}_j(t) = \{1, 2\}$. As a result, the average reward $\tilde{r}_{j,k}(t)$ for arms k in $\tilde{\mathcal{M}}_j(t)$ is 0.5 for arm 1 and 0.4 for arm 2. After sorting the arms in $\tilde{\mathcal{M}}_j(t)$ according to $\tilde{r}_{j,k}(t)$ by descending order and repeating each arm with $\tilde{m}_{j,k}(t)$ times, we have $\tilde{l}_j(t) = \{1, 1, 2\}$.

Afterward, as shown in Line 9-10, player j chooses the element k in $\tilde{l}_j(t)$ with index $i = (t + j) \bmod N + 1$ as a candidate to pull at this round. Moreover, when the player chooses the last element in $\tilde{l}_j(t)$ (*i.e.*, $i = N$), as depicted in Line 13-18, he will explore the arms in $\mathcal{H}_j(t)$ uniformly with probability $1/2$ and the set $\mathcal{H}_j(t)$ is defined as

$$\mathcal{H}_j(t) = \left\{ k' \notin \tilde{\mathcal{M}}_j(t) : \tilde{b}_{j,k'}(t) \geq \tilde{r}_{j,k_a} \right\}. \tag{13}$$

Otherwise, the player will pull arm k as shown in Line 20.

3.3. Theoretical Analysis of SMAA

3.3.1. Performances

Let

$$\delta_0 = \min_{x,y \in \Delta, x \neq y} |x - y|, \tag{14}$$

where $\Delta = \{\mu_k/n : k \in [K], n \in [N]\} \cup \{0\}$ is the set of all possible average rewards. Under Assumption 3.1, we have $\delta_0 > 0$. Now we present the theoretical results about Algorithm 1 on the metrics introduced in Section 2.2.

Theorem 3.3. *Under Assumption 3.1, suppose $0 < \delta < \delta_0/2$. Then when all players follow Algorithm 1, the expected regret for each player j is upper-bounded by*

$$\mathbb{E}\left[\mathrm{Reg}_j(T)\right] \leq \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + 10N^3K(13K + \delta^{-2}). \tag{15}$$

Remark 3.5. The proof is provided in Appendix C.3.2. We highlight that δ_0 and δ is unknown to the algorithm though the bounds depend on δ. This theorem demonstrates that each player can effectively maximize their own reward since the regret is sublinear. In addition, by letting T tend to ∞ and then δ tend to 0, we can get that

$$\limsup_{T \to \infty} \frac{\mathbb{E}\left[\mathrm{Reg}_j(T)\right]}{\log T} \leq \sum_{k \notin \mathcal{M}^*} \frac{z^* - \mu_k}{\mathrm{kl}(\mu_k, z^*)}, \tag{16}$$

which is asymptotically optimal and matches the lower bound in Theorem 3.6 in Section 3.3.2 when reward distributions are Bernoulli.

Remark 3.6. It is also reasonable to define the regret as the gap between each agent's total reward and the average reward in the Nash equilibrium as shown in Theorem 3.1. We demonstrate in Section A.5 that our algorithm still achieves a $O(\log T)$ regret under this formulation.

Theorem 3.4. *Under Assumption 3.1, suppose $0 < \delta < \delta_0/2$. Then when all players follow Algorithm 1, the expected number of non-equilibrium rounds demonstrated in Theorem 3.1 is upper bounded by*

$$\mathbb{E}[\mathrm{NonEqu}(T)] \leq N \sum_{k \notin \mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + 10N^3K(13K + \delta^{-2}). \tag{17}$$

Remark 3.7. The proof is provided in Appendix C.3.3. This theorem demonstrates that when all players follow Algorithm 1, they will eventually converge to the Nash equilibrium since the expected number of non-equilibrium rounds is sublinear. Although it is well known that no internal regret strategies converge in average to the set of correlated equilibria (Nisan et al., 2007), in this paper we prove that our algorithm can converge to the pure Nash equilibrium at each round. As a result, standard no internal regret strategies cannot be adopted directly in our problem setting.

Theorem 3.5. *Under Assumption 3.1, suppose $0 < \delta < \delta_0/2$. Then the policy profile where all players follow Algorithm 1 is an ϵ-Nash equilibrium w.r.t. the game specified by $(\mathcal{S}_{policy}, \{\mathcal{U}_j^{policy}\}_{j=1}^N)$ and is $(\beta, \epsilon + \beta\gamma)$-stable with*

$$\begin{aligned} \beta &= \frac{\delta_0}{z^*}, \\ \epsilon &= \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} \\ &\quad + 10N^3K(13K + \delta^{-2}), \\ \gamma &= \sum_{k \notin \mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + 10N^3K(13K + \delta^{-2}). \end{aligned} \tag{18}$$

Here z^* is provided in Equation (8).

Remark 3.8. The proof is provided in Appendix C.3.4. This theorem demonstrates that any player can not significantly improve his reward by deviation since the policy profile is ϵ-Nash equilibrium with $\epsilon = O(\log T)$. In addition, if any selfish player wants to harm another player's reward significantly (such as $O(T)$), then he will also suffer from a big loss (also $O(T)$). We also note that the bound can be tighter (ϵ, γ can be smaller) in reality since we prove the theorem by assuming the strategic player knows all the parameters beforehand. As a result, SMAA is more stable to the strategic deviation of a single player in reality since the actual ϵ and γ are smaller than those demonstrated in the theorem.

3.3.2. Regret Lower Bound

We now present the regret lower bound for each player. Similar to the previous works on KL-UCB (Garivier & Cappé, 2011; Combes et al., 2015b; Besson & Kaufmann, 2018), we focus on the Bernoulli reward cases. Formally, we use $\mathcal{I}_{N,K}^{\text{ber}} \subseteq \Xi^K$ to denote the set of all problem instances $(\xi_1, \xi_2, \dots, \xi_K)$ that satisfies Assumption 3.1 and each ξ_k is a Bernoulli distribution.

Definition 3.1. We say a policy profile $\boldsymbol{A} = (A_1, A_2, \dots, A_N) \in \mathcal{S}_{\text{policy}}^N$ is *consistent* if for any problem instance $(\xi_1, \xi_2, \dots, \xi_K) \in \mathcal{I}_{N,K}^{\text{ber}}$ with expectations $\mu_1, \mu_2, \dots, \mu_K$ and any $k \in \mathcal{M}^*, j \in [N]$,

$$\forall \alpha \in (0,1], \quad \frac{m_k^*}{N} T - \mathbb{E}[\tau_{j,k}(T)] \le o(T^\alpha). \tag{19}$$

Here $\tau_{j,k}(T)$ is the number of times that arm k is chosen. m_k^* is the number of times arm k is chosen in the equilibrium calculated according to Theorem 3.1. $\mathcal{M}^*$ is the set of arms that in the equilibrium.

We note that Algorithm 1 satisfies the consistent condition since players sequentially choose the arms in the Nash equilibrium and sub-optimal arms are chosen at most $O(\log T)$ times (demonstrated by Lemmas C.3, C.4, and C.5 in Appendix C.3). Any policy profile $\boldsymbol{A}$ that is not consistent cannot achieve a desirable regret guarantee, when $\boldsymbol{A}$ satisfies a fairness condition that all players are expected to pull each arm for similar times. The formal formulation and the proof of this claim can be found in Section A.6. Note that this fairness condition is also mentioned by (Besson & Kaufmann, 2018) and is satisfied by all policy profiles that assigning the same algorithm to all players. Both Algorithm 1 and Algorithm 2 satisify the fairness condition.

With this definition, we show a problem-dependent asymptotic lower bound for consistent algorithms that only use the arms' reward information.

Theorem 3.6. *Suppose a policy profile $\boldsymbol{A} \in \mathcal{S}_{policy}^N$ is consistent as defined in Definition 3.1. Assume $\boldsymbol{A}$ only uses arms' reward information. Formally, for each agent j, $\pi_j(t)$ is $\mathcal{F}_j(t)$-measurable and*

$$\mathcal{F}_j(t) = \sigma(X_{\pi_j(1)}(1), X_{\pi_j(2)}(2), \dots, X_{\pi_j(t-1)}(t-1), U_0, U_1, \dots, U_t). \tag{20}$$

Here $\sigma(\cdot)$ denote the sigma algebra and $U_0, U_1, \dots, U_t$ denote the external sources of randomness to determine $\pi_j(t)$ (e.g., the randomness in Lines 16-17 in Algorithm 1). Then for any problem instance $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_K) \in \mathcal{I}_{N,K}^{\text{ber}}$ with expectations $\mu_1, \mu_2, \dots, \mu_K$, we have

$$\forall k \notin \mathcal{M}^*, \quad \liminf_{T\to\infty} \frac{\mathbb{E}[\tau_{j,k}(T)]}{\log T} \ge \frac{1}{\text{kl}(\mu_k, z^*)}. \tag{21}$$

Furthermore, we have

$$\forall j \in [N], \quad \liminf_{T\to\infty} \frac{\mathbb{E}[\text{Reg}_j(T)]}{\log T} \ge \sum_{k \notin \mathcal{M}^*} \frac{z^* - \mu_k}{\text{kl}(\mu_k, z^*)}. \tag{22}$$

Remark 3.9. The proof is provided in Appendix C.3.5. Here we additionally assume that algorithms only use the arms' reward information to avoid the effect of a strategic player that deliberately collides with other players. Algorithm 1 satisfies the condition. Therefore, under the conditions, our algorithm's regret matches the lower bound asymptotically.

3.4. SMAA without Knowledge of N and Rank

In previous subsections, we assume that players have the knowledge of the total number of players N and each player has a different rank. However, this information may not be available in real-world scenarios. In this case, we further assume that $N \le K$ and there is an indicator to inform each player whether he meets a collision, which is common in previous works (Rosenski et al., 2016; Boursier & Perchet, 2019; Bande & Veeravalli, 2019).

With these assumptions, we could adopt the Musical Chairs approach (Rosenski et al., 2016) to estimate the number of players and assign each player a different rank. The details of the algorithm is provided in Algorithm 2 in Appendix D. We demonstrate that the Musical Chairs approach is robust to a single player's deviation and prove similar results for this algorithm with the conclusions in Section 3.3.1.

Corollary 3.7 (Informal version of Theorem D.2)**.** *Suppose Assumption 3.1 hold. Then Algorithm 2 satisfies that $\forall j \in [N], \mathbb{E}[\text{Reg}_j(T)] \le O(\log T)$. $\mathbb{E}[\text{NonEqu}(T)] \le O(\log T)$. In addition, the policy profile where all players follow Algorithm 2 is an ϵ-Nash equilibrium and is (β, ϵ)-stable with $\epsilon = O(\log T)$ and $\beta = \delta_0/z^*$.*

4. Synthetic Experiments

We validate the effectiveness of the proposed SMAA method through synthetic experiments. The source code is available at `https://github.com/windxrz/SMAA`.

4.1. Experiments When N and Rank Are Known

We first conduct experiments in a setting where players know N and are endowed with different ranks.

Data-generating process In this setting, we set the number of arms as $K = 10$ and the number of players $N \in \{8, 20\}$ to test the models' performance in both $K < N$ and $K \ge N$ scenarios. The number of total rounds T is set to $500,000$. In addition, for each arm's reward distribution, we consider two common distributions supported on $[0,1]$, including the beta distribution and Bernoulli distribution. For the beta distribution $\text{Beta}(\alpha, \beta)$, we randomly sample the

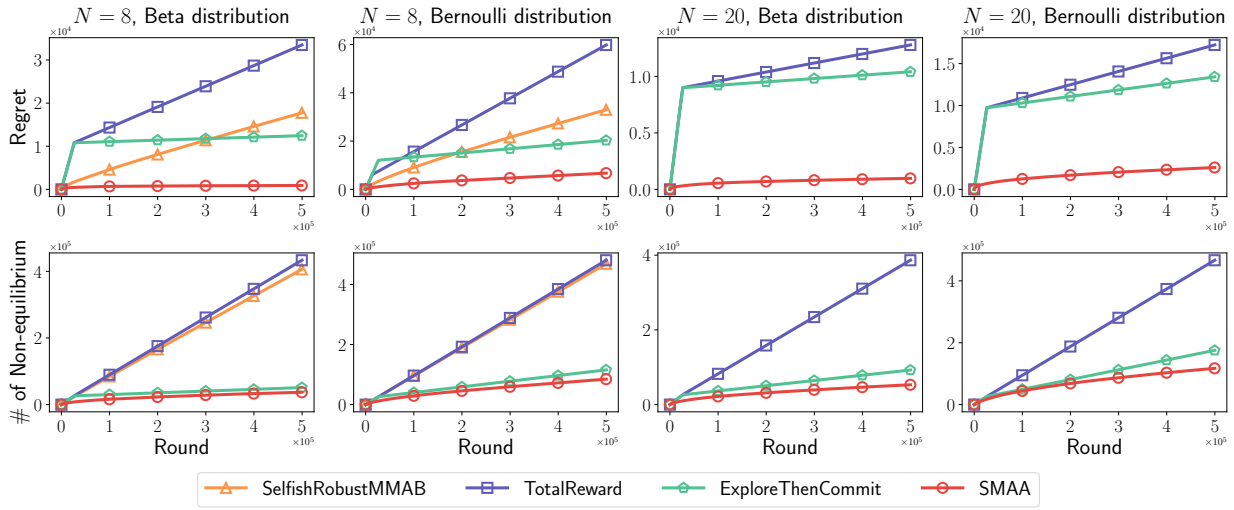

Figure 1. Regret and number of non-equilibrium rounds curve when N and rank is known. K is set to 10. We note that SelfishRobustMMAB (Boursier & Perchet, 2020) considers the non-shareable arm setting and can only be applied to the settings when $N \leq K$.

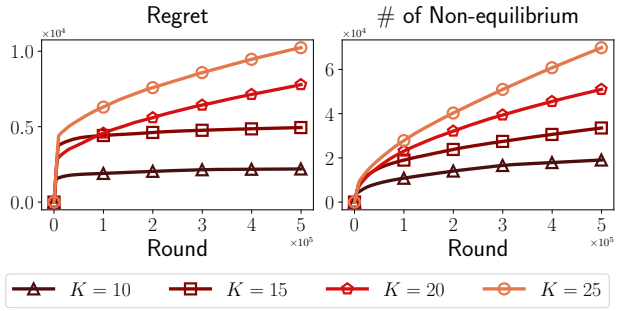

Figure 2. Regret and the number of non-equilibrium rounds curve when N and rank are unknown. Here $N = 8$, $T = 500,000$, and the reward of each arm follows the beta distribution.

two shape parameters α and β uniformly in $[0, 5]$ for each arm. For the Bernoulli distribution $\text{Ber}(p)$, we randomly sample the probability parameter p uniformly in $[0, 1]$.

Baselines We implement the following baselines. **TotalReward** (Bande & Veeravalli, 2019) considers the shareable reward case while its target is to maximize the total reward of all players. **SelfishRobustMMAB** (Boursier & Perchet, 2020) assumes players can be selfish while the arms are not shareable and players get no reward upon collision. **ExploreThenCommit** first randomly explores the rewards of arms and then commits to the Nash equilibrium based on the estimation in the exploration phase. Details of these baselines are provided in Appendix B.

Analysis We implement the experiments for 100 different simulations by resampling the arms' reward distributions. We then evaluate our method (SMAA) and baselines by calculating the average regret among all players and the number of non-equilibrium rounds according to Equations (4) and (5). The results are shown in Figure 1 and our method outperforms all baselines in all settings with various data-generating processes.

On the one hand, the TotalReward and SelfishRobustMMAB baselines consider the different MPMAB scenarios and they can not directly be applied here for the selfish player setting with the averaging collision model. On the other hand, compared with the explore-then-exploit-based baseline ExploreThenCommit, our method shows a better exploration-exploitation trade-off, leading to smaller regret and non-equilibrium rounds.

4.2. Experiments When N and Rank Are Unknown

We then test under the setting when players do not know N and ranks.

Data-generating process In this setting, we fix the number of players as $N = 8$, and the total rounds as $T = 500,000$. In addition, we simulate each arm's reward using the beta distribution. We set $K \in \{10, 15, 20, 25\}$ to test the convergence and performance of our approach.

Analysis Analogous to the experiments when N and rank are known, we implement the experiments for 100 different simulations by resampling the arms' reward distributions. We then evaluate SMAA by calculating the average regret among all players and the number of non-equilibrium rounds. As shown in Figure 2, we observe that our method could effectively converge to Nash equilibrium and achieve small regret since the slopes of these curves become smoother as the number of rounds increases. In addition, we observe from Figure 2 that when the number of arms K becomes larger, the problem becomes more challenging, and both the regrets and the number of non-equilibrium rounds increase. However, our method could still effectively converge since the proportion of non-equilibrium rounds is about 14% $(69,959/500,000)$ when $K = 25$ at round

$500,000$, and the proportion is still decreasing due to the declining slope of the curve.

5. Related Works

Multi-player multi-armed bandit Multi-armed bandit (MAB) problems (Lai et al., 1985; Bubeck et al., 2012; Slivkins et al., 2019; Lattimore & Szepesvári, 2020) have been extensively studied in different online settings (Kveton et al., 2015; Lattimore et al., 2016; Agrawal & Devanur, 2016; Xu et al., 2022; Ferreira et al., 2022). Extending MAB, multi-player multi-armed bandit (MPMAB) (Liu & Zhao, 2008; Jouini et al., 2009; 2010; Brânzei & Peres, 2021) considers the problem where multiple players act on a single multi-armed bandits problem instance (Boursier & Perchet, 2022). Most papers in MPMAB assume that players cannot get any reward when a collision occurs (Rosenski et al., 2016; Besson & Kaufmann, 2018; Boursier & Perchet, 2019; 2020; Shi et al., 2020; 2021; Huang et al., 2022; Lugosi & Mehrabian, 2022). A thorough survey can be found in (Boursier & Perchet, 2022).

Different collision models in MPMAB Several papers assume different reward allocation mechanisms when a collision occurs. Liu et al. (2020; 2021); Jagadeesan et al. (2021); Basu et al. (2021) assume that the player with the maximal expected reward would get the reward. Some works assume a threshold (Bande & Veeravalli, 2019; Youssef et al., 2021; Bande et al., 2021; Magesh & Veeravalli, 2021) or capacity (Wang et al., 2022a;b) for each arm. Shi & Shen (2021); Pacchiano et al. (2021) assume that each player gets a reduced reward when a collision occurs. Boyarski et al. (2021) focus on a heterogeneous reward setting. Our work differs from these works mainly in that we assume an averaging allocation mechanism with homogeneous arms' rewards. In addition, most of these works do not consider selfish player behaviors and the target is to maximize the total welfare.

Strategic players in MPMAB There are two kinds of strategic behaviors in MPMAB, including jammers and selfish players. Jammers (Attar et al., 2012; Wang et al., 2015; Sawant et al., 2018; 2019) aim to perturb the cooperative players as much as possible even at the cost of their rewards. Several works also consider the selfish behaviors in MPMAB while they assume that at most one player will get the reward. Specifically, Boursier & Perchet (2020) considers the standard zero-reward collision model. Liu et al. (2020; 2021); Jagadeesan et al. (2021) consider the stable matching setting. Compared to these works, we assume that arms are shareable.

Game analysis in applications with shareable resources Ranging from algorithm-curated platforms (Hron et al., 2022; Jagadeesan et al., 2022; Ben-Porat et al., 2019a; Yao et al., 2022a;b; 2023) to cognitive radio communications (Riahi & Riahi, 2019; Riahi et al., 2016), game theory plays an important role in allocating limited resources among multiple strategic players. For algorithm-curated platforms, the equilibrium of content providers is analyzed (Basat et al., 2017; Ben-Porat & Tennenholtz, 2018; Ben-Porat et al., 2019a; Jagadeesan et al., 2022; Hron et al., 2022). Regarding cognitive radio communications, *i.e.*, another commonly analyzed strategic scenario, Riahi et al. (2016); Riahi & Riahi (2019) discuss various types of equilibrium among multiple strategic transmitters. However, most of these works analyze the equilibrium only and do not analyze the dynamics of players. Although this gap has been overcome by several recent works (Riahi & Riahi, 2019; Basat et al., 2017; Ben-Porat et al., 2019b; 2020), they still assume that players know the reward for each resource.

In addition, contrary to Yao et al. (2023)'s analysis of the Price of Anarchy (PoA) in coarse correlated equilibria (CCE) for top-K recommendations, our study situates within a MPMAB framework, concentrating on pure Nash equilibrium (PNE). Our aim is to develop algorithms that maximize individual agent revenue and consistently converge to the PNE, diverging from Yao et al. (2023) focus on CCE's PoA and the assumption of no-regret dynamics.

6. Conclusion

In this paper, we propose a novel multi-player multi-armed setting with an averaging allocation model to characterize the selfish behaviors of players in applications with limited and shareable resources. To design policies for each player in this setting, we first analyze the Nash equilibrium when players know the expected reward of each arm. Based on the equilibrium, we further propose a novel Selfish MPMAB with Averaging Allocation (SMAA) approach for each player. We theoretically demonstrate that SMAA could achieve a good regret guarantee for each player when all players follow the algorithm and it is robust to a single player's strategic deviation.

Acknowledgment

Peng Cui's research was supported in part by National Natural Science Foundation of China (No. U1936219, 62141607), and National Key R&D Program of China (No. 2018AAA0102004). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131, 72133002); the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403. We would like to thank Zi Qian and anonymous reviewers for the helpful feedback.

References

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

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

Attar, A., Tang, H., Vasilakos, A. V., Yu, F. R., and Leung, V. C. A survey of security challenges in cognitive radio networks: Solutions and future research directions. *Proceedings of the IEEE*, 100(12):3172–3186, 2012.

Bande, M. and Veeravalli, V. V. Multi-user multi-armed bandits for uncoordinated spectrum access. In *2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 653–657. IEEE, 2019.

Bande, M., Magesh, A., and Veeravalli, V. V. Dynamic spectrum access using stochastic multi-user bandits. *IEEE Wireless Communications Letters*, 10(5):953–956, 2021.

Basat, R. B., Tennenholtz, M., and Kurland, O. A game theoretic analysis of the adversarial retrieval setting. *Journal of Artificial Intelligence Research*, 60:1127–1164, 2017.

Basu, S., Sankararaman, K. A., and Sankararaman, A. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pp. 705–715. PMLR, 2021.

Ben-Porat, O. and Tennenholtz, M. A game-theoretic approach to recommendation systems with strategic content providers. *Advances in Neural Information Processing Systems*, 31, 2018.

Ben-Porat, O., Goren, G., Rosenberg, I., and Tennenholtz, M. From recommendation systems to facility location games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1772–1779, 2019a.

Ben-Porat, O., Rosenberg, I., and Tennenholtz, M. Convergence of learning dynamics in information retrieval games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1780–1787, 2019b.

Ben-Porat, O., Rosenberg, I., and Tennenholtz, M. Content provider dynamics and coordination in recommendation ecosystems. *Advances in Neural Information Processing Systems*, 33:18931–18941, 2020.

Besson, L. and Kaufmann, E. Multi-player bandits revisited. In *Algorithmic Learning Theory*, pp. 56–92. PMLR, 2018.

Boursier, E. and Perchet, V. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Boursier, E. and Perchet, V. Selfish robustness and equilibria in multi-player bandits. In *Conference on Learning Theory*, pp. 530–581. PMLR, 2020.

Boursier, E. and Perchet, V. A survey on multi-player bandits. *arXiv preprint arXiv:2211.16275*, 2022.

Boyarski, T., Leshem, A., and Krishnamurthy, V. Distributed learning in congested environments with partial information. *arXiv preprint arXiv:2103.15901*, 2021.

Brânzei, S. and Peres, Y. Multiplayer bandit learning, from competition to cooperation. In *Conference on Learning Theory*, pp. 679–723. PMLR, 2021.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Combes, R., Magureanu, S., Proutiere, A., and Laroche, C. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 231–244, 2015a.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015b.

Ferreira, K. J., Parthasarathy, S., and Sekar, S. Learning to rank an assortment of products. *Management Science*, 68 (3):1828–1848, 2022.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Haykin, S. Cognitive radio: brain-empowered wireless communications. *IEEE journal on selected areas in communications*, 23(2):201–220, 2005.

Hron, J., Krauth, K., Jordan, M. I., Kilbertus, N., and Dean, S. Modeling content creator incentives on algorithm-curated platforms. *arXiv preprint arXiv:2206.13102*, 2022.

Huang, W., Combes, R., and Trinh, C. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Conference on Learning Theory*, pp. 1990–2012. PMLR, 2022.

Ieong, S., McGrew, R., Nudelman, E., Shoham, Y., and Sun, Q. Fast and compact: A simple class of congestion games. In *AAAI*, volume 5, pp. 489–494, 2005.

Jagadeesan, M., Wei, A., Wang, Y., Jordan, M., and Steinhardt, J. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

Jagadeesan, M., Garg, N., and Steinhardt, J. Supply-side equilibria in recommender systems. *arXiv preprint arXiv:2206.13489*, 2022.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. Multi-armed bandit based policies for cognitive radio's decision making issues. In *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*, pp. 1–6. IEEE, 2009.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. Upper confidence bound based decision making strategies and dynamic spectrum access. In *2010 IEEE International Conference on Communications*, pp. 1–5. IEEE, 2010.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pp. 767–776. PMLR, 2015.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Lattimore, F., Lattimore, T., and Reid, M. D. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Liu, K. and Zhao, Q. A restless bandit formulation of opportunistic access: Indexablity and index policy. In *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pp. 1–5. IEEE, 2008.

Liu, K. and Zhao, Q. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing*, 58(11):5667–5681, 2010.

Liu, L. T., Mania, H., and Jordan, M. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1618–1628. PMLR, 2020.

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22:1–34, 2021.

Lugosi, G. and Mehrabian, A. Multiplayer bandits without observing collision information. *Mathematics of Operations Research*, 47(2):1247–1265, 2022.

Magesh, A. and Veeravalli, V. V. Decentralized heterogeneous multi-player multi-armed bandits with non-zero rewards on collisions. *IEEE Transactions on Information Theory*, 68(4):2622–2634, 2021.

Meshkati, F., Poor, H. V., and Schwartz, S. C. Energy-efficient resource allocation in wireless networks. *IEEE Signal Processing Magazine*, 24(3):58–68, 2007.

Mitola, J. and Maguire, G. Q. Cognitive radio: making software radios more personal. *IEEE personal communications*, 6(4):13–18, 1999.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Pacchiano, A., Bartlett, P., and Jordan, M. I. An instance-dependent analysis for the cooperative multi-player multi-armed bandit. *arXiv preprint arXiv:2111.04873*, 2021.

Riahi, S. and Riahi, A. Game theory for resource sharing in large distributed systems. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(2), 2019.

Riahi, S., El Hore, A., and El Kafi, J. Optimization of resource allocation in wireless systems based on game theory. *International Journal of Computer Sciences and Engineering*, 4(1):01–13, 2016.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

Roughgarden, T. *Selfish routing and the price of anarchy*. MIT press, 2005.

Sawant, S., Hanawal, M. K., Darak, S., and Kumar, R. Distributed learning algorithms for coordination in a cognitive network in presence of jammers. In *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–8. IEEE, 2018.

Sawant, S., Kumar, R., Hanawal, M. K., and Darak, S. J. Learning to coordinate in a decentralized cognitive radio network in presence of jammers. *IEEE Transactions on Mobile Computing*, 19(11):2640–2655, 2019.

Shi, C. and Shen, C. Multi-player multi-armed bandits with collision-dependent reward distributions. *IEEE Transactions on Signal Processing*, 69:4385–4402, 2021.

Shi, C., Xiong, W., Shen, C., and Yang, J. Decentralized multi-player multi-armed bandits with no collision information. In *International Conference on Artificial Intelligence and Statistics*, pp. 1519–1528. PMLR, 2020.

Shi, C., Xiong, W., Shen, C., and Yang, J. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in neural information processing systems*, 34:22392–22404, 2021.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Wang, Q., Ren, K., Ning, P., and Hu, S. Jamming-resistant multiradio multichannel opportunistic spectrum access in cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 65(10):8331–8344, 2015.

Wang, X., Xie, H., and Lui, J. Multi-player multi-armed bandits with finite shareable resources arms: Learning algorithms & applications. *arXiv preprint arXiv:2204.13502*, 2022a.

Wang, X., Xie, H., and Lui, J. C. Multiple-play stochastic bandits with shareable finite-capacity arms. In *International Conference on Machine Learning*, pp. 23181–23212. PMLR, 2022b.

Xu, R., Zhang, X., Li, B., Zhang, Y., Chen, X., and Cui, P. Product ranking for revenue maximization with multiple purchases. In *Advances in Neural Information Processing Systems*, 2022.

Yao, F., Li, C., Nekipelov, D., Wang, H., and Xu, H. Learning from a learning user for optimal recommendations. In *International Conference on Machine Learning*, pp. 25382–25406. PMLR, 2022a.

Yao, F., Li, C., Nekipelov, D., Wang, H., and Xu, H. Learning the optimal recommendation from explorative users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9457–9465, 2022b.

Yao, F., Li, C., Nekipelov, D., Wang, H., and Xu, H. How bad is top-k recommendation under competing content creators? *arXiv preprint arXiv:2302.01971*, 2023.

Youssef, M.-J., Veeravalli, V. V., Farah, J., Nour, C. A., and Douillard, C. Resource allocation in noma-based self-organizing networks using stochastic multi-armed bandits. *IEEE Transactions on Communications*, 69(9): 6003–6017, 2021.

Zhu, B., Karimireddy, S. P., Jiao, J., and Jordan, M. I. Online learning in a creator economy. *arXiv preprint arXiv:2305.11381*, 2023.

A. Additional Theoretical Results

A.1. Justification of Assumption 3.1

Proposition A.1. *Suppose the expected rewards $\mu_1, \mu_2, \ldots, \mu_K \in (0, 1]$ of arms are i.i.d. and drawn from a absolutely continuous probability distribution D with bounded probability densities $p(\mu)$ (i.e., for all $\mu \in (0, 1]$, $p(\mu) \leq M$ for a constant M). Then Assumption 3.1 holds with probability* 1.

The proof is provided in Section C.4.1.

A.2. Extending the Single-Agent Deviation Setting to Multiple-Agent Deviation Setting

We prove that every Nash equilibrium in Theorem 3.1 is a strong Nash equilibrium.

Proposition A.2. *For a strategy profile $\boldsymbol{s}$, we say strategies $\boldsymbol{s}'_B \in \prod_{j\in B} \mathcal{S}_{single}$ are a beneficial deviation for a subset B of players if*

$$\forall j \in B, \quad \mathcal{U}_j^{single}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) \geq \mathcal{U}_j^{single}(\boldsymbol{s}) \tag{23}$$

with the inequality holding strictly for at least one player of B. Then every Nash equilibrium $\boldsymbol{s}$ in Theorem 3.1 has no coalition of players with a beneficial deviation.

The proof is provided in Section C.4.2.

A.3. Randomized Strategy at Each Round

At each round, the strategy set for each player is the set of all arms, *i.e.*, $\mathcal{S}_{\text{single}} = [K]$. Now suppose player j follows randomized strategy $\boldsymbol{\sigma}_j = (\sigma_{j,1}, \sigma_{j,2}, \ldots, \sigma_{j,K})$ on $\mathcal{S}_{\text{single}}$ where $\sigma_{j,k}$ represents the probability for player j to choose arm k. Let $\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 \times \boldsymbol{\sigma}_2 \times \cdots \times \boldsymbol{\sigma}_N$ be the strategy profile of all players. Then the social welfare under a mixed strategy is given by

$$W(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{s}\sim\boldsymbol{\sigma}}\left[\sum_{k=1}^{K} \mu_k \mathbb{I}[m_k(\boldsymbol{s}) > 0]\right] = \sum_{k=1}^{K} \mu_k \mathbb{P}_{\boldsymbol{s}\sim\boldsymbol{\sigma}}[m_k(\boldsymbol{s}) > 0]. \tag{24}$$

The welfare under the pure Nash equilibrium Theorem 3.1 is given by

$$W^{\text{PNE}} = \sum_{k\in\mathcal{M}^*} \mu_k. \tag{25}$$

Definition A.1 (Mixed Nash equilibrium)**.** The randomized strategy profile $\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 \times \boldsymbol{\sigma}_2 \times \cdots \times \boldsymbol{\sigma}_N$ is a mixed Nash equilibrium if for every player $j \in [N]$ and every unilateral deviation s'_j, we have

$$\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}[\mathcal{U}_j(\boldsymbol{s})] \geq \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}[\mathcal{U}_j(s'_j, \boldsymbol{s}_{-j})]. \tag{26}$$

We provide the following property to show that any symmetric mixed Nash equilibrium's total welfare is not larger than the welfare obtained by the pure Nash equilibrium at each round.

Proposition A.3. *Suppose $\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 \times \boldsymbol{\sigma}_2 \times \cdots \times \boldsymbol{\sigma}_N$ is a symmetric mixed Nash equilibrium (i.e., $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = \cdots = \boldsymbol{\sigma}_N$). Then*

$$W(\boldsymbol{\sigma}) \leq W^{PNE}. \tag{27}$$

The proof is provided in Section C.4.3. We further find that it is actually possible that players pull arms that is not in $\mathcal{M}^*$ when all players follow a symmetric mixed Nash equilibrium. However, the welfare under the symmetric mixed Nash equilibrium is not larger than the welfare under the pure Nash equilibrium as shown by Proposition A.3.

Example A.1. Set $N = K = 3$ and $\mu_1, \mu_2, \mu_3 = 1, 0.6, 0.48$. By simulation, we have the profile $\boldsymbol{\sigma}$ such that

$$\forall j \in [K], \quad \sigma_{j,1} \approx 0.705, \sigma_{j,2} \approx 0.254, \sigma_{j,3} \approx 0.041 \tag{28}$$

formulates a symmetric mixed Nash equilibrium. We can verify that all arms are possible to be chosen in the mixed Nash equilibrium while the social welfare is given by

$$W(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\sum_{j=1}^{N} \mathcal{U}_j^{\text{single}}(\boldsymbol{s})\right] \approx 1.382. \tag{29}$$

In addition, we have $\mathcal{M}^* = \{1, 2\}$ and the welfare for the pure Nash equilibrium is

$$W^{\text{PNE}} = 1 + 0.6 = 1.6. \tag{30}$$

A.4. Formal Analysis of the bound on PoA

Let $\mathcal{C}^*$ be the set of the top $\min\{N, K\}$ arms with the highest expected rewards. Then the total welfare under the equilibrium demonstrated in Theorem 3.1 is given by

$$W^{\text{PNE}} = \sum_{k \in \mathcal{M}^*} \mu_k \tag{31}$$

while the total welfare for maximal social welfare case is given by

$$W^{\text{MAX}} = \sum_{k \in \mathcal{C}^*} \mu_k. \tag{32}$$

PoA is measured as follows

$$\text{PoA} = W^{\text{MAX}} / W^{\text{PNE}}. \tag{33}$$

Proposition A.4 (Restatement of Proposition 3.2)**.** *PoA is bounded between* 1 *and* $(N + \min\{K, N\} - 1)/N$*. Furthermore, PoA can take value at* 1 *and any close to* $(N + \min\{K, N\} - 1)/N$*.*

The proof of Proposition 3.2 can be found in Section C.2. We further analyze the effect of N and K on the upper bound.

- The upper bound increases *w.r.t.* K when $K < N$ while remains unchanged when $K \geq N$. This is because only at most N arms with highest expected rewards are in the set $\mathcal{C}^*$ while remaining arms do not affect the value of PoA.
- The upper bound increases *w.r.t.* N when $N < K$ and decreases when $N \geq K$. This is because when $N < K$, the difference between W^{MAX} and W^{PNE} can be $(N-1)/N \cdot W^{\text{PNE}}$ and this difference increases *w.r.t.* N. In addition, when $N \geq K$, it is more easier to make all arms in the equilibrium set $\mathcal{M}^*$ when N is large. Even when some arms in $\mathcal{C}^*$ are not in the equilibrium, the number of such arms are bounded by K, making them not significantly affect the PoA value.

A.5. Another Regret Formulation and Regret Bound Analysis

Consider an alternative regret that compares each agent's total reward with the average reward in the Nash equilibrium. Formally, it is given by the follows

$$\text{Reg}'_j(T) = \sum_{t=1}^{T} \left(r^* - R_j(t)\right), \tag{34}$$

where

$$r^* = \frac{1}{N} \sum_{k \in \mathcal{M}^*} \mu_k. \tag{35}$$

We can provide the regret bound *w.r.t.* this regret by the following proposition.

Proposition A.5. *Under Assumption 3.1, suppose* $0 < \delta < \delta_0/2$*. Then when all players follow Algorithm 1, the expected regret for each player* j *is upper-bounded by*

$$\mathbb{E}\left[\text{Reg}'_j(T)\right] \leq \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\text{kl}(\mu_k + \delta, z^* - \delta)} + O(1). \tag{36}$$

The proof is provided in Section C.4.4.

A.6. Justification of Definition 3.1

Proposition A.6. *Suppose a policy profile $\boldsymbol{A} \in \mathcal{S}_{policy}^N$ satisfies*

1. ***Fairness condition.*** $|\mathbb{E}[\tau_{j,k}(T)] - \mathbb{E}[\tau_{j',k}(T)]| \leq o(T^\alpha)$ *for all* $\alpha > 0$, $j, j' \in [N]$ *and* $k \in [K]$.
2. ***No regret condition.*** $\mathbb{E}[\mathrm{Reg}_j(T)] \leq o(T^\alpha)$ *for all* $j \in [N]$ *and* $\alpha > 0$.

Then $\boldsymbol{A}$ *is consistent as defined in Definition 3.1.*

The proof is provided in Section C.4.5.

B. Experimental Details

B.1. Implementation details of different methods

- **SMAA (Ours)**. We use a hyper-parameter β to control the strength between exploration and exploitation. Specifically, the KL index in Equation (10) is modified as

$$\hat{b}_{j,k}(t) = \sup\left\{q \geq \hat{\mu}_{j,k}(t) : \tau_{j,k}(t)\,\mathrm{kl}(\hat{\mu}_{j,k}(t), q) \leq \beta \cdot f(t)\right\}. \tag{37}$$

  We search the hyper-parameter $\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

- **SMAA (Ours) when N and rank are unknown**. Besides the hyper-parameter β in the exploration-exploitation phase, we introduce another hyper-parameter for this method in the Musical Chairs phase. Specifically, we set the parameter T_0 in Algorithm 3 as $\eta \cdot \left\lceil 50K^2 \log(4T)\right\rceil$ and η is searched in $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

- **SelfishRobustMMAB** (Boursier & Perchet, 2020). This baseline considers the setting where no player would get the reward if collisions occur. As a result, it could only apply to the scenarios $K \geq N$. It provides algorithms for each player that is robust to the deviation of a single selfish player under this setting. This method also adopts the KL-UCB to calculate the index of sub-optimal arms, as shown in Equation (37). We also search the hyper-parameter $\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

- **TotalReward** (Bande et al., 2021). This baseline considers the setting where players can get the reward when collisions occur. However, the target of this baseline is to maximize the total reward of all players. When $K \geq N$, the optimal solution at each round is to make all players pull the best N arms. When $K \leq N$, all arms should be pulled in each round. To further reduce the regrets of each player, after assigning k players to each arm, we make the rest $N - K$ players follow the Nash equilibrium in each round. The whole method is conducted in an explore-then-commit fashion. As a result, we set the algorithm to randomly explore the arms at the first $\alpha \log T$ rounds. Afterward, the algorithm commits to the optimal solution at each round. We search the hyper-parameter $\alpha \in \{100, 200, 500, 1000, 2000\}$.

- **ExploreThenCommit**. We further implement an explore-then-commit-based method. Specifically, we first adopt an exploration stage where all players pull the arms randomly. After the exploration stage, each player estimates the expected reward of each arm and computes the Nash equilibrium *w.r.t.* the estimations. They then commit to the estimated equilibrium till the last round. The exploration phase has $\alpha \log T$ rounds where α is a hyper-parameter searched in $\{100, 200, 500, 1000, 2000\}$.

C. Omitted Proofs

C.1. Proof of Theorem 3.1

Proof. (1) We first show that there exists a strategy profile $\boldsymbol{s}$ such that $\boldsymbol{m}(\boldsymbol{s}) = \boldsymbol{m}^*$.

We only need to prove that $\sum_{k=1}^K m_k^* = N$. Note that z^* exists since $h(z) \to \infty$ when $z \to 0$ and $h(z) = 0$ when $z > 1$. Moreover, under Assumption 3.1, for any $\delta < \delta_0$ (δ_0 is defined in Equation (14)), $h(z) - h(z + \delta) \leq 1$. As a result, there must exist $z \in \mathbb{R}^+$ such that $h(z) = N$. Therefore, $h(z^*) = N$ since $h(z)$ is right continuous. Now we can get that $\sum_{k=1}^K m_k^* = h(z^*) = N$.

(2) We then show that any strategy profile $\boldsymbol{s}$ that satisfies $\boldsymbol{m}(\boldsymbol{s}) = \boldsymbol{m}^*$ is a Nash equilibrium.

Suppose a player j deviate from the strategy profile and change the chosen arm from $a = s_j$ to b ($a \neq b$). By definition, we have $m_a^* > 0$. In addition, since $m_b^* = \lfloor \mu_b / z^* \rfloor$, we have $m_b^* + 1 > \mu_b / z^*$. As a result, the reward after deviation is $\mu_b/(m_b^* + 1) < z^*$. However, the original reward is $\mu_a / n_a^* \geq z^*$, which means that the deviation only leads to the decrease in the reward of the player. Therefore, $\boldsymbol{m}^*$ is a Nash equilibrium of the game.

We note that we have demonstrated the existence of the Nash equilibrium through points (1) and (2).

(3) Afterward, we show that for any strategy profile $\boldsymbol{s}$ that is a Nash equilibrium, $\boldsymbol{m}(\boldsymbol{s}) = \boldsymbol{m}^*$.

Under Assumption 3.1, suppose there exists a Nash equilibrium $\boldsymbol{s}$ such that $\boldsymbol{m}(\boldsymbol{s}) \neq \boldsymbol{m}^*$. Let $\boldsymbol{m}' = \boldsymbol{m}(\boldsymbol{s})$. Then there exist two arms a and b such that $\boldsymbol{m}_a^* > \boldsymbol{m}_a'$ and $\boldsymbol{m}_b^* < \boldsymbol{m}_b'$. Using similar techniques with the first part, it is easy to show that a player deviate from arm b to arm a in the profile $\boldsymbol{s}$ will lead to an increase in the reward, which results in a contradiction.

Now the claim follows. □

C.2. Proofs of Proposition 3.2

Proof. (1) Lower bound 1. The PoA is naturally lower bounded by 1 and can take on this value when the expected rewards of the top $\min(N, K)$ arms are very close, which is possible for any values of N and K.

(2) Upper bound $(N + \min\{K, N\} - 1)/N$. Due to the definition of z^* in Theorem 3.1, we have $W^{\text{PNE}}/N \geq z^* > \mu_k$ for any $k \notin \mathcal{M}^*$. As a result, for any arm $k \in \mathcal{C}^* \backslash \mathcal{M}^*$, $\mu_k < W^{\text{PNE}}/N$. In addition, due to the property of Nash equilibrium demonstrated in Theorem 3.1, all arms in the equilibrium have the highest expected rewards. As a result, $\mathcal{M}^* \subseteq \mathcal{C}^*$. Hence,

$$W^{\text{MAX}} = W^{\text{PNE}} + \sum_{k \in \mathcal{C}^* \backslash \mathcal{M}^*} \mu_k \leq W^{\text{PNE}} + |\mathcal{C}^* \backslash \mathcal{M}^*| W^{\text{PNE}}/N \leq W^{\text{PNE}} \left(1 + (\min(N, K) - 1)/N\right). \tag{38}$$

As a result,

$$\text{PoA} = W^{\text{MAX}}/W^{\text{PNE}} \leq (N + \min\{K, N\} - 1)/N. \tag{39}$$

In addition, the exact value of PoA can be taken any close to $(N + \min\{K, N\} - 1)/N$ by setting $\mu_1 = 1$ and the rewards of other arms very close to but smaller than $1/N$. □

C.3. Proofs in Section 3.3

C.3.1. Basic Notations and Lemmas

Fix δ such that $0 < \delta < \delta_0/2$. Let $K' = N \lceil K/N \rceil$. Define the following random sets

$$\begin{aligned}
\mathcal{A}_j &= \left\{ t > K' : \mathcal{M}^* \backslash \tilde{\mathcal{M}}_j(t) \neq \emptyset \right\}, \\
\mathcal{B}_j &= \left\{ t > K' : \exists k \in \tilde{\mathcal{M}}_j(t), |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta \right\}, \\
\mathcal{C}_j &= \left\{ t > K' : \exists k \in \mathcal{M}^*, \tilde{b}_{j,k}(t) < \mu_k \right\}, \\
\mathcal{D}_j &= \left\{ t \in \mathcal{A}_j \backslash (\mathcal{B}_j \cup \mathcal{C}_j) : \exists k \in \mathcal{M}^* \backslash \tilde{\mathcal{M}}_j(t), |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta \right\}, \\
\mathcal{J} &= \{ t > K' : \exists j \in [K], t \in (\mathcal{A}_j \cup \mathcal{B}_j) \}.
\end{aligned} \tag{40}$$

Let l^* be the optimal list of the arm chosen in a block *w.r.t.* the expected rewards $\{\mu_k\}_{k=1}^K$, using the similar procedures with that in Equation (12). Formally, we sort the arms in the Nash equilibrium $\mathcal{M}^*$ according to $r_k^* = \mu_k^*/m_k^*$ by descending order and get the indices of the arms $k_1^*, k_2^*, \ldots, k_{|\mathcal{M}^*|}^*$. Then we sequentially choose these arms and each arm k_i^* are further chosen by m_i^* times. The arms list l^* is given by

$$l^* = \left(\underbrace{k_1^*, \ldots, k_1^*}_{m_{k_1^*}^* \text{ times}}, \ldots, \underbrace{k_{|\mathcal{M}^*|}^*, \cdots, k_{|\mathcal{M}^*|}^*}_{m_{k_{|\mathcal{M}^*|}^*}^* \text{ times}} \right). \tag{41}$$

Lemma C.1. *Under Assumption 3.1, we have*

$$z^* = \min_{k\in[K], m_k^*>0} \mu_k/m_k^*, \tag{42}$$

where z^* is given by Equation (8).

Lemma C.2. *Under Assumption 3.1, suppose $t > K'$ and $t \notin (\mathcal{A}_j \cup \mathcal{B}_j)$. Let k_0 be the last element in $\tilde{l}_j(t)$. Then $l^* = \tilde{l}_j(t)$, and $|\tilde{r}_{j,k_0}(t) - z^*| \le \delta$ where z^* is given by Equation (8).*

Remark C.1. The proof is provided in Appendix C.3.7. Lemma C.2 demonstrates that when $t \notin (\mathcal{A}_j \cup \mathcal{B}_j)$, player j could find the Nash equilibrium of the game as shown in Theorem 3.1 and choose arms according to l^* in Equation (41).

Lemma C.3. *Under Assumption 3.1, we have $\mathcal{A}_j \cup \mathcal{B}_j \subseteq \mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j$.*

Lemma C.4. *Under Assumption 3.1, we have $\mathbb{E}[|\mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j|] \le 8N^2K(12K + \delta^{-2})$.*

Lemma C.5. *Under Assumption 3.1, define*

$$\mathcal{G}_{j,k} \triangleq \{K' < t \le T : t \notin (\mathcal{A}_j \cup \mathcal{B}_j), \pi_j(t) = k\} \tag{43}$$

for any $k \notin \mathcal{M}^*$. Then

$$\mathbb{E}[|\mathcal{G}_{j,k}|] \le \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + 5 + 2\delta^{-2}. \tag{44}$$

Remark C.2. The proofs of Lemma C.3, Lemma C.4, and Lemma C.5 follow the techniques in (Wang et al., 2020) and details are provided in Appendix C.3.8, Appendix C.3.9, and Appendix C.3.10, respectively. We highlight that we consider a more complex setting since players need to calculate the Nash equilibrium in each round according to Theorem 3.1. As a result, the details of the proofs become more complicated.

C.3.2. Proof of Theorem 3.3

Proof. Consider the set $\mathcal{J}$ defined in Equation (40), according to Lemma C.3 and Lemma C.4, we have

$$\mathbb{E}[|\mathcal{J}|] \le \sum_{j=1}^{N} \mathbb{E}[|\mathcal{A}_j \cup \mathcal{B}_j|] \le \sum_{j=1}^{N} \mathbb{E}[|\mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j|] \le 8N^3K(12K + \delta^{-2}). \tag{45}$$

As a result, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_j(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\le K' + \mathbb{E}\left[\sum_{t=K'+1}^{T} \mathbb{I}[t \in \mathcal{J}]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=K'+1}^{T} \mathbb{I}[t \notin \mathcal{J}]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\le K' + \mathbb{E}[|\mathcal{J}|] + \mathbb{E}\left[\sum_{t=K'+1}^{T} \mathbb{I}[t \notin \mathcal{J}]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\le 8N^3K(13K + \delta^{-2}) + \mathbb{E}\left[\sum_{t=K'+1}^{T} \mathbb{I}[t \notin \mathcal{J}]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - \mu_{\pi_j(t)}\right)\right].
\end{aligned} \tag{46}$$

When $t > K'$ and $t \notin J$, according to Lemma C.2, each player j could find the correct Nash equilibrium $\mathcal{M}^*$ and calculate the correct arms list l^*. Let k_0 be the last element in l^*. As a result, at most one player do not follow the equilibrium $\mathcal{M}^*$ by the construction of Algorithm 1. Formally, this indicates that $\forall k \in \mathcal{M}^* \backslash \{k_0\}$, $M_k(t) = m_k^*$ and $M_{k_0}(t) \ge m_{k_0}^* - 1$. As a result, when $\pi_j(t) \in \mathcal{M}^*$, it will not increase the reward if deviation by the property of the equlibrium and the choice of k_0, *i.e.*,

$$\mathbb{I}[t \notin \mathcal{J}, \pi_j(t) \in \mathcal{M}^*]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - \mu_{\pi_j(t)}]\right) = 0. \tag{47}$$

In addition, when $\pi_j(t) \notin \mathcal{M}^*$, because $\forall k \in \mathcal{M}^* \backslash \{k_0\}$, $M_k(t) = m_k^*$ and $M_{k_0}(t) \geq m_{k_0}^* - 1$, it will maximize the reward if deviating to arm k_0. As a result, $\max_{k\in[k]} \mu_k/(M_k(t) + \mathbb{I}[\pi_j(t) \neq k]) = z^*$ in this case. Therefore, according to Lemma C.5 and the definition of $\mathcal{G}_{j,k}$ in Equation (43), we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_j(T)] &\leq 8N^3K(13K + \delta^{-2}) + \mathbb{E}\left[\sum_{k\notin\mathcal{M}^*}\sum_{t=K'+1}^{T} \mathbb{I}[t \notin \mathcal{J}, t \in \mathcal{G}_{j,k}]\left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - \mu_{\pi_j(t)}\right)\right] \\
&= 8N^3K(13K + \delta^{-2}) + \mathbb{E}\left[\sum_{k\notin\mathcal{M}^*}\sum_{t=K'+1}^{T} \mathbb{I}[t \notin \mathcal{J}, t \in \mathcal{G}_{j,k}]\,(z^* - \mu_k)\right] \\
&\leq 8N^3K(13K + \delta^{-2}) + \mathbb{E}\left[\sum_{k\notin\mathcal{M}^*}(z^* - \mu_k)\sum_{t=K'+1}^{T} \mathbb{I}[t \in \mathcal{G}_{j,k}]\right] \\
&\leq 8N^3K(13K + \delta^{-2}) + \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (5 + 2\delta^{-2})K \\
&\leq 10N^3K(13K + \delta^{-2}) + \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)}.
\end{aligned} \tag{48}$$

Now the claim follows. □

C.3.3. Proof of Theorem 3.4

Proof. Let k_0 be the last element in l^*. Similar to the proof of Theorem 3.3 in Appendix C.3.2, we have $\mathbb{E}[\mathcal{J}] \leq 8N^3K(12K + \delta^{-2})$. In addition, when $t > K'$ and $t \notin J$, we have $\forall k \in \mathcal{M}^* \backslash \{k_0\}$, $M_k(t) = m_k^*$ and $M_{k_0}(t) \geq m_{k_0}^* - 1$. As a result, if players do not achieve the Nash equilibrium at round t, there must exist a player j such that $\pi_j(t) \notin \mathcal{M}^*$. As a result, according to Lemma C.5 and the definition of $\mathcal{G}_{j,k}$ in Equation (43), we have

$$\begin{aligned}
\mathbb{E}[\mathrm{NonEqu}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left[\exists k \in [K], M_k(t) \neq m_k^*\right]\right] \\
&\leq K' + \mathbb{E}\left[\sum_{t=K'+1}^{T} \mathbb{I}\left[\exists k \in [K], M_k(t) \neq m_k^*, t \in \mathcal{J}\right]\right] + \mathbb{E}\left[\sum_{t=K'}^{T} \mathbb{I}\left[\exists k \in [K], M_k(t) \neq m_k^*, t \notin \mathcal{J}\right]\right] \\
&\leq K' + \mathbb{E}[|\mathcal{J}|] + \mathbb{E}\left[\sum_{t=K'}^{T} \mathbb{I}\left[\exists j \in [N], \pi_j(t) \notin \mathcal{M}^*, t \notin \mathcal{J}\right]\right] \\
&\leq 8N^3K(13K + \delta^{-2}) + \sum_{j=1}^{N}\sum_{k\notin\mathcal{M}^*} \mathbb{E}\left[\sum_{t=K'}^{T} \mathbb{I}[t \notin \mathcal{J}, \pi_j(t) = k]\right] \\
&\leq 8N^3K(13K + \delta^{-2}) + \sum_{j=1}^{N}\sum_{k\notin\mathcal{M}^*} \mathbb{E}[|\mathcal{G}_{j,k}|] \\
&\leq 8N^3K(13K + \delta^{-2}) + N\sum_{k\notin\mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (5 + 2\delta^{-2})NK \\
&\leq 10N^3K(13K + \delta^{-2}) + N\sum_{k\notin\mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)}.
\end{aligned} \tag{49}$$

Now the claim follows. □

C.3.4. Proof of Theorem 3.5

Proof. For the ϵ-Nash equilibrium part, we note that Algorithm 1 only uses the information of $X_{\pi_j(t)}(t)$, which means that the behaviors of strategic players will not affect the policy of any other player. As a result, the reward of a strategic player

deviate from Algorithm 1 to other algorithm is at most the regret demonstrated in Theorem 3.3. Hence, the profile where all players follow Algorithm 1 is an ϵ-Nash equilibrium with

$$\epsilon = \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + 10N^3K(13K + \delta^{-2}). \tag{50}$$

Consider the $(\beta, \epsilon + \beta\gamma)$-stable part. Suppose player j is strategic and player i follows Algorithm 1. When $t \in \mathcal{J}$ or $t \leq K'$, the increase of player j's reward is at most 1 and the decrease in player i's reward is at most 1. Let k_0 be the last element in l^*.

We first show that when $t > K'$ and $t \notin \mathcal{J}$, the loss of player i's reward in each round due to the deviation of player j is at most z^*. When $t > K'$ and $t \notin \mathcal{J}$, according to the proof of Theorem 3.3 in Appendix C.3.2, we have $\forall k \in \mathcal{M}^* \backslash \{k_0\}$, $M_k(t) = m_k^*$ and $M_{k_0}(t) \geq m_{k_0}^* - 1$. We consider different cases of the values of $\pi_i(t)$

- Consider the case when $\pi_i(t) \in \mathcal{M}^* \backslash \{k_0\}$. Then if $\pi_i(t) = \pi_j(t)$, player j can not lead to a decrease in player i' reward. If $\pi_i(t) \neq \pi_j(t)$, player j will deviate to pull arm $\pi_i(t)$ to make player i worse. In this case, when $\pi_i(t) \neq k_0$, the loss of player i's reward is

$$\frac{\mu_k}{m_k^*} - \frac{\mu_k}{m_k^* + 1} = \frac{\mu_k}{m_k^*(m_k^* + 1)} \leq \frac{\mu_k}{m_k^* + 1} < z^*. \tag{51}$$

- Consider the case when $\pi_i(t) = k_0$. If $M_{k_0}(t) = m_{k_0}^* - 1$, then $m_{k_0}^* > 1$ and the loss of player i's reward is

$$\frac{\mu_{k_0}}{m_{k_0}^* - 1} - \frac{\mu_{k_0}}{m_{k_0}^*} = \frac{\mu_{k_0}}{m_{k_0}^*(m_{k_0}^* - 1)} \leq \frac{\mu_{k_0}}{m_{k_0}^*} = z^*. \tag{52}$$

- Consider the case when $\pi_i(t) \notin \mathcal{M}^*$. Then the loss of player i's reward is $\mu_{\pi_i(t)}/2 < z^*$.

We then consider the change of player j's reward by deviation when $t > K'$ and $t \notin \mathcal{J}$. When $\pi_j(t) \notin \mathcal{M}^*$, player j could increase the reward by at most $z^* - \mu_{\pi_j(t)}$. When $\pi_j(t) \in \mathcal{M}^*$, player j will also incur a loss at least δ_0.

Suppose the player j deviate to policy s', let $\mathcal{Q}$ be the set of rounds that player j choose to deviate from Algorithm 1. Let

$$\begin{aligned} &\mathcal{Q}_1 = \{t \in \mathcal{Q} : t \in \mathcal{J} \cup \{1, 2, \ldots, K'\}\}, \mathcal{Q}_2 = \{t \in \mathcal{Q} : t \notin \mathcal{J}, t > K', \pi_j(t) \notin \mathcal{M}^*\}, \\ &\mathcal{Q}_3 = \{t \in \mathcal{Q} : t \notin \mathcal{J}, t > K', \pi_j(t) \in \mathcal{M}^*\}. \end{aligned} \tag{53}$$

Then $\mathcal{Q}_1$, $\mathcal{Q}_2$, and $\mathcal{Q}_3$ are disjoint and $\mathcal{Q} = \mathcal{Q}_1 \cup \mathcal{Q}_2 \cup \mathcal{Q}_3$. Then

$$\begin{aligned} &\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j}) \geq \mathrm{Rew}_i(T; \boldsymbol{s}) - |\mathcal{Q}_1| - z^*(|\mathcal{Q}_2| + |\mathcal{Q}_3|), \\ &\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j}) \leq \mathrm{Rew}_j(T; \boldsymbol{s}) + |\mathcal{Q}_1| + \left(\sum_{k \notin \mathcal{M}^*} (z^* - \mu_k) |\{t \in \mathcal{Q}_2 : \pi_j(t) = k\}| \right) - \delta_0 |\mathcal{Q}_3|. \end{aligned} \tag{54}$$

Taking expectations, according to Lemma C.5 and the definition of $\mathcal{G}_{j,k}$ in Equation (43), we have

$$\begin{aligned} &\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_i(T; \boldsymbol{s})] \\ \geq &- \mathbb{E}[|\mathcal{Q}_1|] - z^* \mathbb{E}[|\mathcal{Q}_2|] - z^* \mathbb{E}[|\mathcal{Q}_3|] \\ \geq &- 8N^3K(13K + \delta^{-2}) - z^* \sum_{k \notin \mathcal{M}^*} \mathbb{E}[|\mathcal{G}_{j,k}|] - z^* \mathbb{E}[|\mathcal{Q}_3|] \\ \geq &- 8N^3K(13K + \delta^{-2}) - z^* \left(K(5 + 2\delta^{-2}) + \sum_{k \notin \mathcal{M}^*} \frac{\log T + \log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} \right) - z^* \mathbb{E}[|\mathcal{Q}_3|] \\ \geq &- 10N^3K(13K + \delta^{-2}) - \sum_{k \notin \mathcal{M}^*} \frac{\log T + \log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} - z^* \mathbb{E}[|\mathcal{Q}_3|] \end{aligned} \tag{55}$$

In addition,

$$\begin{aligned}
&\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_j(T; \boldsymbol{s})] \\
\leq& \mathbb{E}[|\mathcal{Q}_1|] + \left(\sum_{k \notin \mathcal{M}^*} (z^* - \mu_k) \mathbb{E}[|\{t \in \mathcal{Q}_2 : \pi_j(t) = k\}|] \right) - \delta_0 \mathbb{E}[|\mathcal{Q}_3|] \\
\leq& \mathbb{E}[|\mathcal{Q}_1|] + \left(\sum_{k \notin \mathcal{M}^*} (z^* - \mu_k) \mathbb{E}[|\mathcal{G}_{j,k}|] \right) - \delta_0 \mathbb{E}[|\mathcal{Q}_3|] \\
\leq& 10N^3K(13K + \delta^{-2}) + \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} - \delta_0 \mathbb{E}[|\mathcal{Q}_3|].
\end{aligned} \tag{56}$$

Here the last inequality is due to Equation (48). Because

$$\gamma = 10N^3K(13K + \delta^{-2}) + \sum_{k \notin \mathcal{M}^*} \frac{\log T + \log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)}, \tag{57}$$

we have

$$\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_i(T; \boldsymbol{s})] \geq -\gamma - z^* \mathbb{E}[|\mathcal{Q}_3|] \quad \text{and} \quad \mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_j(T; \boldsymbol{s})] \leq \epsilon - \delta_0 \mathbb{E}[|\mathcal{Q}_3|]. \tag{58}$$

As a result, for any $u \in \mathbb{R}^+$, if $\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_i(T; \boldsymbol{s})] \leq -u$, we have $\mathbb{E}[|\mathcal{Q}_3|] \geq (u - \gamma)/z^*$. Then

$$\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_j(T; \boldsymbol{s})] \leq \epsilon - \delta_0 \frac{u - \gamma}{z^*} = \epsilon + \frac{\delta_0 \gamma}{z^*} - \frac{\delta_0}{z^*} u. \tag{59}$$

Now the claim follows. □

C.3.5. Proof of Theorem 3.6

Proof. The proof follows the techniques of the proof of Lemma 3 in (Liu & Zhao, 2010) and Theorem 1 in (Anantharam et al., 1987).

Let any problem instance $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K) \in \mathcal{I}_{N,K}^{\mathrm{ber}}$. Here $\xi_1, \xi_2, \ldots, \xi_K$ are Bernoulli distributions with expectations $\mu_1, \mu_2, \ldots, \mu_K$. z^* and $\boldsymbol{m}^*$ are calculated according to Theorem 3.1. $\mathcal{M}^* = \{k \in [K] : m_k^* > 0\}$. δ_0 is given by Equation (14). Consider any sub-optimal arm k such that $\mu_k < z^*$. As a result, $m_k^* = 0$.

For any $\beta \in (0, 1)$, choose a parameter λ such that

$$z^* < \lambda < z^* + \delta_0, \quad |\mathrm{kl}(\mu_k, \lambda) - \mathrm{kl}(\mu_k, z^*)| \leq \beta\, \mathrm{kl}(\mu_k, z^*), \quad \text{and} \quad \boldsymbol{\xi}' \triangleq (\xi_1, \xi_2, \ldots, \xi_{k-1}, \xi_k', \xi_{k+1}, \ldots, \xi_K) \in \Xi. \tag{60}$$

Here ξ_k' is the Bernoulli distribution with expectation λ. We use z'^*, $\boldsymbol{m}'^*$, and $\mathcal{M}'^*$ to denote the Nash equilibrium *w.r.t.* the problem instance $\boldsymbol{\xi}'$. We use $\mathbb{P}_{\boldsymbol{\xi}}[E]$ and $\mathbb{P}_{\boldsymbol{\xi}'}[E]$ to denote the probability of an event E when the problem instance is $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, respectively.

According to Theorem 3.1 and Lemma C.1, let k_0 be the arm in $\mathcal{M}^*$ such that $\mu_{k_0}/m_{k_0}^* = z^*$. Since $z^* < \lambda < z^* + \delta$, we have that

$$\forall k' \in \mathcal{M}^* \backslash \{k_0\}, \quad m_{k'}^* \leq \frac{\mu_{k'}}{\left\lfloor \frac{\mu_{k'}}{m_{k'}^*} \right\rfloor} < \frac{\mu_{k'}}{z^* + \delta} < \frac{\mu_{k'}}{z^* + \delta} < \frac{\mu_{k'}}{\lambda} < \frac{\mu_{k'}}{z^*} < m_{k'}^* + 1. \tag{61}$$

As a result, $\lfloor \mu_{k'}/\lambda \rfloor = m_{k'}^*$. In addition, $\lfloor \mu_{k_0}/\lambda \rfloor = m_{k_0}^* - 1$. As a result,

$$\sum_{k'=1}^{K} \left\lfloor \frac{\mu_{k'}'}{\lambda} \right\rfloor = \left(\sum_{k' \in \mathcal{M}^*} \left\lfloor \frac{\mu_{k'}'}{\lambda} \right\rfloor \right) + \left\lfloor \frac{\lambda}{\lambda} \right\rfloor + \left(\sum_{k' \notin \mathcal{M}^* \cup \{k\}} \left\lfloor \frac{\mu_{k'}'}{\lambda} \right\rfloor \right) = N - 1 + 1 + 0 = N. \tag{62}$$

and we have $m_k'^* = 1$. Therefore, since algorithm A is consistent, we have $T/N - \mathbb{E}_{\boldsymbol{\xi}'}[\tau_{j,k}(T)] \leq o(T^\alpha)$ for any $\alpha \in (0, \beta)$. Define a random function $c(T)$ as

$$c(T) \triangleq \max\left\{ \mathbb{E}_{\boldsymbol{\xi}'} \left[\frac{T}{N} - \tau_{j,k}(T) \right], 0 \right\} - \frac{T}{N} + \tau_{j,k}(T). \tag{63}$$

We have $\mathbb{E}_{\boldsymbol{\xi}'}[c(T)] \geq 0$ and

$$\mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + \frac{T}{N} - \tau_{j,k}(T)\right] = o(T^{\alpha}). \tag{64}$$

Because $c(T) + T/N - \tau_{j,k}(T) \geq 0$ a.e., we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + \frac{T}{N} - \tau_{j,k}(T)\right] =& \mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + \frac{T}{N} - \tau_{j,k}(T)\middle|\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right]\mathbb{P}_{\boldsymbol{\xi}'}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] \\
&+ \mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + \frac{T}{N} - \tau_{j,k}(T)\middle|\tau_{j,k}(T) \geq \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right]\mathbb{P}_{\boldsymbol{\xi}'}\left[\tau_{j,k}(T) \geq \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] \\
\geq& \mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + \frac{T}{N} - \tau_{j,k}(T)\middle|\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right]\mathbb{P}_{\boldsymbol{\xi}'}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] \\
\geq& \left(\mathbb{E}_{\boldsymbol{\xi}'}[c(T)] + \frac{T}{N} - \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right)\mathbb{P}_{\boldsymbol{\xi}'}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right].
\end{aligned} \tag{65}$$

Let $S_{k,1}, S_{k,2}, \ldots$ be the independent observations from arm k. Define

$$L_t \triangleq \sum_{i=1}^{t} \log\left(\frac{\mathbb{P}_{\boldsymbol{\xi}}[X_k = S_{k,i}]}{\mathbb{P}_{\boldsymbol{\xi}'}[X_k = S_{k,i}]}\right) \tag{66}$$

and event

$$C \triangleq \left\{\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}, L_{\tau_{j,k}(T)} \leq \left(1 - \frac{\alpha+\beta}{2}\right)\log T\right\}. \tag{67}$$

Combining Equations (64) and (65), we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\xi}'}[C] \leq \mathbb{P}_{\boldsymbol{\xi}'}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] &\leq \frac{\mathbb{E}_{\boldsymbol{\xi}'}\left[c(T) + T/N - \tau_{j,k}(T)\right]}{\mathbb{E}_{\boldsymbol{\xi}'}[c(T)] + T/N - (1-\beta)\log T/\mathrm{kl}(\mu_k,\lambda)} \\
&\leq \frac{o(T^{\alpha})}{T/N - (1-\beta)\log T/\mathrm{kl}(\mu_k,\lambda)} = o(T^{\alpha-1}).
\end{aligned} \tag{68}$$

Let $C_s \triangleq \{\tau_{j,k}(T) = s, L_s \leq (1-(\alpha+\beta)/2)\log T\}$. We have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\xi}'}[C_s] &= \int_{\{\tau_{j,k}(T)=s, L_s \leq (1-(\alpha+\beta)/2)\log T\}} \mathrm{d}\mathbb{P}_{\boldsymbol{\xi}'} = \int_{\{\tau_{j,k}(T)=s, L_s \leq (1-(\alpha+\beta)/2)\log T\}} \prod_{i=1}^{s} \frac{\mathbb{P}_{\boldsymbol{\xi}'}[X_k = S_{k,i}]}{\mathbb{P}_{\boldsymbol{\xi}}[X_k = S_{k,i}]} \mathrm{d}\mathbb{P}_{\boldsymbol{\xi}} \\
&\geq \int_{\{\tau_{j,k}(T)=s, L_s \leq (1-(\alpha+\beta)/2)\log T\}} \exp\left(-(1-(\alpha+\beta)/2)\log T\right) \mathrm{d}\mathbb{P}_{\boldsymbol{\xi}} = T^{(\alpha+\beta)/2-1}\mathbb{P}_{\boldsymbol{\xi}}[C_s].
\end{aligned} \tag{69}$$

Since C_s $(1 \leq s < (1-\beta)\log T/\mathrm{kl}(\mu_k,\lambda))$ are disjoint, we have

$$\mathbb{P}_{\boldsymbol{\xi}}[C] \leq T^{1-(\alpha+\beta)/2}\mathbb{P}_{\boldsymbol{\xi}'}[C] = o\left(T^{\frac{\alpha-\beta}{2}}\right) \to 0, \quad \text{when } T \to \infty. \tag{70}$$

In addition, by the law of large numbers, we have that under $\mathbb{P}_{\boldsymbol{\xi}}$, when $t \to \infty$, $L_t/t \to \mathrm{kl}(\mu_k,\lambda) > 0$. As a result, under $\mathbb{P}_{\boldsymbol{\xi}}$, when $t \to \infty$, $\max_{i \leq t} L_i/t \to \mathrm{kl}(\mu_k,\lambda) > 0$. Therefore, since $\beta > \alpha$, we have $(1-(\alpha+\beta)/2) > (1-\beta)$ and

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\xi}}\left[L_i > (1-(\alpha+\beta)/2)\log T \text{ for some } i < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] = 0. \tag{71}$$

As a result,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\xi}}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}, L_{\tau_{j,k}(T)} > \left(1 - \frac{\alpha+\beta}{2}\right)\log T\right] = 0. \tag{72}$$

Combining Equations (70) and (72), we have

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\xi}}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{\mathrm{kl}(\mu_k,\lambda)}\right] = 0 \tag{73}$$

According to Equation (60), we have $\mathrm{kl}(\mu_k, \lambda) \geq (1-\beta)\,\mathrm{kl}(\mu_k, z^*)$. Therefore,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\xi}}\left[\tau_{j,k}(T) < \frac{(1-\beta)\log T}{(1+\beta)\,\mathrm{kl}(\mu_k, z^*)}\right] = 0. \tag{74}$$

Now the claim for $\tau_{j,k}(T)$ follows by letting $\beta \to 0$.

For the regret part, since for each arm k that is not in the Nash equilibrium $\mathcal{M}^*$, the regret for deviation is at least $z^* - \mu_k$. Therefore, we have

$$\begin{aligned}
\mathrm{Reg}_j(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - \sum_{t=1}^{T} R_j(t)\right] \\
&\geq \sum_{k\notin\mathcal{M}^*} \mathbb{E}\left[\sum_{t\in[T],\pi_j(t)=k} \left(\max_{k\in[k]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\geq \sum_{k\notin\mathcal{M}^*} \mathbb{E}[\tau_{j,k}(T)](z^* - \mu_k).
\end{aligned} \tag{75}$$

Now the claim for $\mathrm{Reg}_j(T)$ follows by applying Equation (21). □

C.3.6. Proof of Lemma C.1

Proof. On the one hand, for any k such that $m_k^* > 0$, since $m_k^* = \lfloor \mu_k/z^* \rfloor$, we have $m_k^* \leq \mu_k/z^*$ and $z^* \leq \mu_k/m_k^*$. As a result, $z^* \leq \min_{k\in[K], m_k^*>0} \mu_k/m_k^*$. On the other hand, let $z = \min_{k\in[K], m_k^*>0} \mu_k/m_k^*$. For any for any k such that $m_k^* > 0$, we have that $\mu_k/m_k^* \geq z$. Hence, we have $\mu_k/z \geq m_k^*$. As a result,

$$h(z) = \sum_{k=1}^{K} \left\lfloor \frac{\mu_k}{z} \right\rfloor \geq \sum_{k=1}^{K} m_k^* = N. \tag{76}$$

By the definition of z^*, we have $z^* \geq z$. To conclude, we have $z^* \geq \min_{k\in[K], m_k^*>0} \mu_k/m_k^*$ and $z^* \leq \min_{k\in[K], m_k^*>0} \mu_k/m_k^*$. □

C.3.7. Proof of Lemma C.2

Proof. Let

$$\tilde{z}_j(t) = \sup\left\{z > 0 : \tilde{h}_{j,t}(z) \triangleq \sum_{k=1}^{K} \left\lfloor \frac{\tilde{\mu}_{j,k}(t)}{z} \right\rfloor \geq N\right\} \tag{77}$$

and we have $\tilde{z}_j(t) = \tilde{r}_{j,k}(t)$ according to Theorem 3.1 and Lemma C.1.

(1) We first show that $\tilde{l}_j(t) = l^*$.

Since $t \notin \mathcal{A}_j$, we have $\mathcal{M}^* \subseteq \tilde{\mathcal{M}}_j(t)$. Let $z_0 = z^* - \delta$. On the one hand, Fix any $k \in \tilde{\mathcal{M}}_j(t)$. Under Assumption 3.1, according to Theorem 3.1, we know that $z^* > \mu_k/(m_k^* + 1)$ and there exists $k' \in \mathcal{M}^*$ such that $z^* = \mu_{k'}/m_{k'}^*$. By the choice of δ, we have

$$z^* > \frac{\mu_k}{m_k^* + 1} + 2\delta. \tag{78}$$

As a result, since $t \notin \mathcal{B}_j$, we have

$$\frac{\tilde{\mu}_{j,k}(t)}{z_0} < \frac{\mu_k + \delta}{z^* - \delta} < \frac{\mu_k + \delta}{\frac{\mu_k}{m_k^*+1} + \delta} = (m_k^* + 1) \cdot \frac{\mu_k + \delta}{\mu_k + (m_k^* + 1)\delta} < (m_k^* + 1). \tag{79}$$

On the other hand, according to Theorem 3.1, we have $z^* \leq \mu_k/m_k^*$. As a result,

$$\frac{\tilde{\mu}_{j,k}(t)}{z_0} > \frac{\mu_k - \delta}{z^* - \delta} > \frac{\mu_k - \delta}{\frac{\mu_k}{m_k^*} - \delta} = m_k^* \cdot \frac{\mu_k - \delta}{\mu_k - m_k^*\delta} > m_k^*. \tag{80}$$

Combining the above two equations and we get for any $k \in \tilde{\mathcal{M}}_j(t)$,

$$\left\lfloor \frac{\tilde{\mu}_{j,k}(t)}{z_0} \right\rfloor = m_k^*. \tag{81}$$

As a result, $\tilde{h}_{j,t}(z_0) \geq \sum_{k \in \mathcal{M}^*} m_k^* = N$. Therefore, $\tilde{z}_j(t) \geq z_0$, and

$$\tilde{m}_{j,k}(t) = \left\lfloor \frac{\tilde{\mu}_{j,k}(t)}{\tilde{z}_{j,k}(t)} \right\rfloor \leq \left\lfloor \frac{\tilde{\mu}_{j,k}(t)}{z_0^*} \right\rfloor = m_k^*, \quad \forall k \in \tilde{\mathcal{M}}_j(t). \tag{82}$$

Since $m_k^* = 0$ for all $k \in \tilde{\mathcal{M}}_j(t) \backslash \mathcal{M}$, we have $\tilde{\mathcal{M}}_j(t) = \mathcal{M}^*$ and $\forall k \in \mathcal{M}^*$, $\lfloor \tilde{\mu}_{j,k}(t)/z_0 \rfloor = \lfloor \tilde{\mu}_{j,k}(t)/\tilde{z}_j(t) \rfloor$. Now we can conclude that $\tilde{\boldsymbol{m}}_j^*(t) = \boldsymbol{m}^*$.

Furthermore, let $k, k' \in \mathcal{M}^*$ ($k \neq k'$) and suppose $\mu_k/m_k^* < r_k < r_{k'} = \mu_{k'}/m_{k'}^*$. Then by the choice of δ, we have $\mu_k'/m_{k'}^* - \mu_k/m_k^* > 2\delta$. As a result, since $t \notin \mathcal{B}_j$, we have

$$\tilde{r}_{j,k}(t) - \tilde{r}_{j,k'}(t) = \frac{\tilde{\mu}_{j,k}(t)}{\tilde{m}_{j,k}(t)} - \frac{\tilde{\mu}_{j,k'}(t)}{\tilde{m}_{j,k'}(t)} = \frac{\tilde{\mu}_{j,k}(t)}{m_k^*} - \frac{\tilde{\mu}_{j,k'}(t)}{m_{k'}^*} \geq \frac{\mu_k - \delta}{m_k^*} - \frac{\mu_{k'} + \delta}{m_{k'}^*} > \left(2 - \frac{1}{m_k^*} - \frac{1}{m_{k'}^*}\right)\delta > 0. \tag{83}$$

This equation indicates that we could get the same order if sorting $\{r_k^* = \mu_k/m_k^*\}_{k \in \mathcal{M}^*}$ and $\{\tilde{r}_{j,k}(t)\}_{k \in \mathcal{M}^*}$ by descending order. Together with the conclusion $\tilde{\boldsymbol{m}}_j(t) = \boldsymbol{m}^*$, we can get $\tilde{l}_j(t) = l^*$.

(2) We now show that $|\tilde{r}_{j,k_0}(t) - z^*| \leq \delta$ where k_0 is the last element in $\tilde{l}_j(t)$.

From the proof of part (1), we know that $z^* - \delta \leq \tilde{z}_j(t)$. In addition, since $t \notin \mathcal{B}_j$, we have

$$\tilde{z}_j(t) = \tilde{r}_{j,k_0}(t) = \frac{\tilde{\mu}_{j,k_0}(t)}{\tilde{m}_{j,k_0}(t)} < \frac{\mu_{k_0} + \delta}{m_{k_0}^*} < \frac{\mu_{k_0}}{m_{k_0}^*} + \delta = z^* + \delta. \tag{84}$$

Here the last equation is based on the fact that $l^* = \tilde{l}(t)$ and hence $z^* = \min_{k' \in \mathcal{M}^*} \mu_{k'}/m_{k'}^* = \mu_{k_0}/m_{k_0}^*$. Now the claim follows. □

C.3.8. Proof of Lemma C.3

Proof. We only need to show that $\mathcal{A}_j \subseteq \mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j$. Let $t \in \mathcal{A}_j \backslash (\mathcal{B}_j \cup \mathcal{C}_j)$.

Because $t \in \mathcal{A}_j$, we have that there exists $k \in \mathcal{M}^* \backslash \tilde{\mathcal{M}}_j(t)$. Since

$$\left\{\mathcal{M}^* \backslash \tilde{\mathcal{M}}_j(t) \neq \emptyset\right\} = \left\{\tilde{\mathcal{M}}_j(t) \backslash \mathcal{M}^* \neq \emptyset\right\} \cup \left\{\tilde{\mathcal{M}}_j(t) \subsetneq \mathcal{M}^*\right\}. \tag{85}$$

We consider these two cases respectively.

(1) Consider the $\mathcal{M}^* \backslash \tilde{\mathcal{M}}_j(t) \neq \emptyset$ case.

Let $k' \in \tilde{\mathcal{M}}_j(t) \backslash \mathcal{M}^*$. Therefore, we have $\tilde{\mu}_{j,k}(t) \leq \tilde{\mu}_{j,k'}(t)$. Otherwise, a player deviate from arm k' to arm k will increase the reward, which indicates that $\tilde{\boldsymbol{m}}_j(t)$ is not a Nash equilibrium *w.r.t.* $\{\tilde{\mu}_{j,k}(t)\}_{k=1}^K$ and leads to a contradiction.

In addition, since $t \notin \mathcal{B}_j$, we have that $|\tilde{\mu}_{j,k'}(t) - \mu_{k'}| < \delta$. Furthermore, by the definition of δ_0 and the property of the Nash equilibrium $\boldsymbol{m}^*$ *w.r.t.* $\{\mu_k\}_{k=1}^K$, we have

$$\frac{\mu_k}{m_k^*} - \mu_{k'} \geq \delta_0 > 2\delta. \tag{86}$$

As a result,

$$\mu_k - \tilde{\mu}_{j,k}(t) \geq m_k^*(\mu_{k'} + 2\delta) - \tilde{\mu}_{j,k'}(t) \geq 2\delta + (\mu_{k'} - \tilde{\mu}_{j,k'}(t)) > \delta. \tag{87}$$

Therefore, $t \in \mathcal{D}_j$.

(2) Consider the $\tilde{\mathcal{M}}_j(t) \subsetneq \mathcal{M}^*$ case.

Since $\tilde{\mathcal{M}}_j(t) \subsetneq \mathcal{M}^*$, there exists $k' \in \tilde{\mathcal{M}}_j(t)$ such that $\tilde{m}_{j,k'}(t) > m^*_{k'}$. Because $k \notin \tilde{\mathcal{M}}_j(t)$, we have $\tilde{\mu}_{j,k}(t) \leq \tilde{\mu}_{j,k'}(t)/\tilde{m}_{j,k'}(t)$. Furthermore, by the definition of δ_0 and the property of the Nash equilibrium $\boldsymbol{m}^*$ *w.r.t.* $\{\mu_k\}_{k=1}^K$, we have

$$\mu_k \geq \frac{\mu_{k'}}{m^*_{\mu_{k'}}+1} + 2\delta \geq \frac{\mu_{k'}}{\tilde{m}_{j,k'}(t)} + 2\delta. \tag{88}$$

In addition, since $t \notin \mathcal{B}_j$, we have that $|\tilde{\mu}_{j,k'}(t) - \mu_{k'}| < \delta$. As a result,

$$\mu_k - \tilde{\mu}_{j,k}(t) \geq \frac{\mu_{k'}}{\tilde{m}_{j,k'}(t)} + 2\delta - \tilde{\mu}_{j,k}(t) > \frac{\tilde{\mu}_{j,k'}(t) - \delta}{\tilde{m}_{j,k'}(t)} + 2\delta - \tilde{\mu}_{j,k}(t) \geq \delta\left(2 - \frac{1}{\tilde{m}_{j,k'}(t)}\right) \geq \delta. \tag{89}$$

Therefore, $t \in \mathcal{D}_j$.

Now the claim follows. □

C.3.9. Proof of Lemma C.4

Proof. Since $\mathbb{E}[|\mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j|] \leq \mathbb{E}[|\mathcal{B}_j|] + \mathbb{E}[|\mathcal{C}_j|] + \mathbb{E}[|\mathcal{D}_j|]$, we provide the upper bounds of $\mathbb{E}[|\mathcal{B}_j|]$, $\mathbb{E}[|\mathcal{C}_j|]$, and $\mathbb{E}[|\mathcal{D}_j|]$, respectively. Let $\mathcal{F}_t$ be the σ-algebra generated by $\{X_k(s), \forall k \in [K]\}_{s=1}^t$.

(1) We first show that $\mathbb{E}[|\mathcal{B}_j|] \leq NK(17 + 4\delta^{-2})$.

For any $k \in [K]$, let $\mathcal{B}_{j,k} = \left\{t > K' : k \in \tilde{\mathcal{M}}_j(t), |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta\right\}$. Let $1 \leq p_k(t) \leq N$ be the index of the last element that is k in $\tilde{l}_t$. Define

$$\mathcal{B}_{j,k,1} = \{t \in \mathcal{B}_{j,k} : t \equiv p_k(t) (\operatorname{mod} N)\}. \tag{90}$$

Let $H = \{t > K' : t \equiv p_k(t) (\operatorname{mod} N)\}$ and $C_t = \mathbb{I}[\pi_j(t) = k]$. Note that $\mathbb{I}[t \in H]$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{P}[C_t = 1 | t \in H] \geq 1/2$ by the design of the algorithm. According to Lemma E.1, we can get that

$$\mathbb{E}[|\mathcal{B}_{j,k,1}|] \leq \sum_{t>K'} \mathbb{P}\left[t \in H, |\hat{\mu}_{j,k}(t) - \mu_k| \geq \delta\right] \leq 4(4 + \delta^{-2}). \tag{91}$$

$\forall t \in H$, it is always the last possible time to choose arm k in the corresponding block w.r.t. t. As a result, $\forall t \in H$, $\hat{\mu}_{j,k}(t)$ is the estimations adopted in the next block. Formally, $\forall s > N$ and $s \in \mathcal{B}_{j,k}$, we have $\tilde{\mu}_{j,k}(s) = \hat{\mu}_{j,k}(\max\{t \in H : t < s\})$ and $\max\{t \in H : t < s\} \in \mathcal{B}_{j,k,1}$. In addition, each $t \in \mathcal{B}_{j,k,1}$ corresponds to at most N elements in $\mathcal{B}_{j,k}$, which means that

$$\mathbb{E}[|\mathcal{B}_{j,k}|] \leq N + N \cdot \mathbb{E}[|\mathcal{B}_{j,k,1}|] \leq N(17 + 4\delta^{-2}). \tag{92}$$

Now by the union bound, we have

$$\mathbb{E}[|\mathcal{B}_j|] \leq \sum_{k=1}^K \mathbb{E}[|\mathcal{B}_{j,k}|] \leq NK(17 + 4\delta^{-2}). \tag{93}$$

(2) We then show that $\mathbb{E}[|\mathcal{C}_j|] \leq 60N^2$.

For any $k \in \mathcal{M}^*$, let $\mathcal{C}_{j,k} = \left\{t > K' : \tilde{b}_{j,k}(t) < \mu_k\right\}$. By the design of our algorithm, $\tilde{b}_{j,k}(t)$ is updated at the start of each block. As a result

$$\mathbb{E}[|\mathcal{C}_{j,k}|] = \sum_{t>K'} \mathbb{P}\left[\tilde{b}_{j,k}(t) < \mu_k\right] \leq N \sum_{t\geq 0} \mathbb{P}\left[\tilde{b}_{j,k}(tN+1) < \mu_k\right] \leq N + N \sum_{t\geq 1} \mathbb{P}\left[\hat{b}_{j,k}(tN) < \mu_k\right]. \tag{94}$$

According to Lemma E.2 (Theorem 10 in (Garivier & Cappé, 2011)), we can get that for all $t \geq 3$,

$$\mathbb{P}\left[\hat{b}_{j,k}(tN) < \mu_k\right] \leq \lceil f(tN) \log f(tN) \rceil e^{1-f(tN)}. \tag{95}$$

Here $f(tN) = \log(tN) + 4\log(\log(tN))$. As a result,

$$\begin{aligned}\mathbb{E}[|\mathcal{C}_{j,k}|] &\leq 3N + N\sum_{t\geq 3}\lceil f(tN)\log f(tN)\rceil e^{1-f(tN)} \\ &\leq 3N + N\sum_{t\geq 3}\lceil f(s)\log f(s)\rceil e^{1-f(s)} \leq 60N,\end{aligned} \tag{96}$$

where the last inequality is based on the proof of Lemma 4 in (Wang et al., 2020). Therefore, by the union bound, we have

$$\mathbb{E}[|\mathcal{C}_j|] \leq \sum_{k\in\mathcal{M}^*}\mathbb{E}[|C_{j,k}|] \leq 60N^2. \tag{97}$$

(3) Finally, we show that $\mathbb{E}[|\mathcal{D}_j|] \leq N^2K(17K + 4\delta^{-2})$.

For any $k \in \mathcal{M}^*$, let

$$\mathcal{D}_{j,k} = \left\{t \in \mathcal{A}_j\backslash(\mathcal{B}_j \cup \mathcal{C}_j) : k \notin \tilde{\mathcal{M}}_j(t) \text{ and } |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta\right\}. \tag{98}$$

Fix $k \in \mathcal{M}^*$. Suppose $t \in \mathcal{D}_{j,k}$. Since $t \notin C_j$, we have $\tilde{b}_{j,k}(t) \geq \mu_k$. In addition, since $t \notin \mathcal{B}_j$, we have that for any $k' \in \tilde{\mathcal{M}}_j(t)$, $|\tilde{\mu}_{j,k'}(t) - \mu_{k'}| < \delta$. Let k_0 be the last element in this list $\tilde{l}_j(t)$.

In addition, because $t \in \mathcal{A}_j$, we have $\mathcal{M}^*\backslash\tilde{\mathcal{M}}_j(t) \neq \emptyset$. Because $\left\{\mathcal{M}^*\backslash\tilde{\mathcal{M}}_j(t) \neq \emptyset\right\} = \left\{\tilde{\mathcal{M}}_j(t)\backslash\mathcal{M}^* \neq \emptyset\right\} \cup \left\{\tilde{\mathcal{M}}_j(t) \subsetneq \mathcal{M}^*\right\}$. We show that $k \in \mathcal{H}_j(t)$ in these two cases, respectively.

(a) Suppose $\tilde{\mathcal{M}}_j(t)\backslash\mathcal{M}^* \neq \emptyset$. Let k' be any element in $\tilde{\mathcal{M}}_j(t)\backslash\mathcal{M}^*$. As a result, since $k \in \mathcal{M}^*$ and $k' \notin \mathcal{M}^*$, under Assumption 3.1, we have $\mu_k \geq \mu_{k'} + 2\delta$. In addition, by the construction of $\tilde{l}_j(t)$, we have $\tilde{\mu}_{j,k'}(t) \geq \tilde{r}_{j,k'}(t) \geq \tilde{r}_{j,k_0}(t)$. As a result,

$$\tilde{b}_{j,k}(t) \geq \mu_k \geq \mu_{k'} + 2\delta \geq \tilde{\mu}_{j,k'}(t) + \delta > \tilde{r}_{j,k_0}(t). \tag{99}$$

As a result, $k \in \mathcal{H}_j(t)$ by the construction of the algorithm.

(b) Suppose $\tilde{\mathcal{M}}_j(t) \subsetneq \mathcal{M}^*$. There exists $k' \in \tilde{\mathcal{M}}_j(t)$ such that $\tilde{m}_{j,k'}(t) > m^*_{k'}$. Because $k \notin \tilde{\mathcal{M}}_j(t)$, we have $\tilde{\mu}_{j,k}(t) \leq \tilde{\mu}_{j,k'}(t)/\tilde{m}_{j,k'}(t)$. Furthermore, by the definition of δ_0 and the property of the Nash equilibrium $\boldsymbol{m}^*$ *w.r.t.* $\{\mu_k\}_{k=1}^K$, we have

$$\mu_k \geq \frac{\mu_{k'}}{m^*_{\mu_{k'}} + 1} + 2\delta \geq \frac{\mu_{k'}}{\tilde{m}_{j,k'}(t)} + 2\delta. \tag{100}$$

As a result, by the construction of $\tilde{l}_j(t)$, we have $\tilde{\mu}_{j,k'}(t)/\tilde{m}_{j,k'}(t) \geq \tilde{r}_{j,k_0}(t)$. Therefore, we can get

$$\tilde{b}_{j,k}(t) \geq \mu_k \geq \frac{\mu_{k'}}{\tilde{m}_{j,k'}(t)} + 2\delta \geq \frac{\tilde{\mu}_{j,k'}(t) - \delta}{\tilde{m}_{j,k'}(t)} + 2\delta \geq \tilde{r}_{j,k_0}(t) + \delta\left(2 - \frac{1}{\tilde{m}_{j,k'}(t)}\right) > \tilde{r}_{j,k_0}(t). \tag{101}$$

To conclude, from part (a) and (b), we get that $k \in \mathcal{H}_j(t)$. Therefore, when $t_0 = N$ in Algorithm 1, arm k will be chosen with probability at least $1/(2K)$. Now use the similar techniques in the proof of part (1). Let

$$\mathcal{D}_{j,k,1} = \{t \in \mathcal{D}_{j,k} : (t + j) \equiv N - 1 (\operatorname{mod} N)\}. \tag{102}$$

By the construction of the algorithm, for each $t \in \mathcal{D}_{j,k,1}$, arm k will be chosen with probability at least $1/(2K)$. Let $H = \{t \in \mathcal{A}_j\backslash(\mathcal{B}_j \cup \mathcal{C}_j) : k \notin \tilde{\mathcal{M}}_j(t) \text{ and } (t + j) \equiv N - 1 (\operatorname{mod} N)\}$, $C_t = \mathbb{I}[\pi_j(t) = k]$. According to Lemma E.1, we have

$$\mathbb{E}[|\mathcal{D}_{j,k,1}|] = \sum_{t\geq 1}\mathbb{P}[t \in \mathcal{H}, |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta] \leq 16K^2 + 4K\delta^{-2}. \tag{103}$$

Similar to the proof of part (1), we have that

$$\mathbb{E}[|\mathcal{D}_{j,k}|] \leq N(1 + \mathbb{E}[|\mathcal{D}_{j,k,1}|]) \leq NK(17K + 4\delta^{-2}). \tag{104}$$

Finally, by the union bound, we have

$$\mathbb{E}[|\mathcal{D}_j|] \leq \sum_{k\in\mathcal{M}^*}\mathbb{E}[|\mathcal{D}_{j,k}|] \leq N^2K(17K + 4\delta^{-2}). \tag{105}$$

Now the claim follows by combining Equations (93), (97), and (105). □

C.3.10. Proof of Lemma C.5

Proof. Define $w(t) = \sum_{s=1}^{t} \mathbb{I}[s \in \mathcal{G}_{j,k}]$ and $s_0 = (\log T + 4\log(\log T))/\operatorname{kl}(\mu_k + \delta, z^* - \delta)$ where z^* is provided in Equation (8). Further, let

$$\mathcal{G}_{j,k,1} = \{t \in \mathcal{G}_{j,k} : |\tilde{\mu}_{j,k}(t) - \mu_k| \geq \delta\} \quad \text{and} \quad \mathcal{G}_{j,k,2} = \{t \in \mathcal{G}_{j,k} : w(t) < s_0 + 1\}. \tag{106}$$

(1) We first show that $\mathcal{G}_{j,k} \subseteq (\mathcal{G}_{j,k,1} \cup \mathcal{G}_{j,k,2})$.

Suppose $\mathcal{G}_{j,k} \backslash (\mathcal{G}_{j,k,1} \cup \mathcal{G}_{j,k,2}) \neq \emptyset$. Let $t \in \mathcal{G}_{j,k} \backslash (\mathcal{G}_{j,k,1} \cup \mathcal{G}_{j,k,2})$. Since $t \in \mathcal{G}_{j,k}$, we have that $t \notin \mathcal{A}_j$ and $\mathcal{M}^* = \tilde{\mathcal{M}}_j(t)$. As a result, $k \notin \tilde{\mathcal{M}}_j(t)$. Therefore, by the design of Algorithm 1, arm k will be chosen at most one time in the block corresponding to round t. As a result, $\tilde{\tau}_{j,k}(t) \geq w(t) - 1$. In addition, since $t \notin \mathcal{G}_{j,k,2}$, we have $w(t) \geq s_0 + 1$. Hence, we have

$$\tilde{\tau}_{j,k}(t) \geq s_0. \tag{107}$$

Let $k' = \tilde{l}_{j,N}(t) \in \mathcal{M}^*$, *i.e.*, the last element in this list $\tilde{l}_j(t)$ with the smallest estimated average reward $\tilde{r}_{j,k'}(t)$. We note that $\pi_j(t) = k$ only when $\tilde{b}_{j,k}(t) \geq \tilde{r}_{j,k'}(t)$ by the choice of $\mathcal{H}_j(t)$ in Line 9 of Algorithm 1. In addition, since $t \notin (\mathcal{A}_j \cup \mathcal{B}_j)$, according to Lemma C.2, we have that

$$\tilde{b}_{j,k}(t) \geq \tilde{r}_{j,k'}(t) \geq z^* - \delta. \tag{108}$$

Furthermore, since $t \notin \mathcal{G}_{j,k,1}$, we have $|\tilde{\mu}_{j,k}(t) - \mu_k| < \delta$. Because $k \notin \mathcal{M}^*$, under Assumption 3.1, we have $\mu_k < z^*$ (Otherwise, a player that chooses arm $q = k^*_{|\mathcal{M}^*|}$ deviate to choose arm k will increase the reward). As a result, by the construction of δ, we have $z^* - \mu_k > 2\delta$. Hence,

$$\tilde{\mu}_{j,k}(t) \leq \mu_k + \delta < z^* - \delta. \tag{109}$$

As a result, we can get that

$$s_0 \operatorname{kl}(\tilde{\mu}_{j,k}(t), z^* - \delta) \leq \tilde{\tau}_{j,k}(t) \operatorname{kl}(\tilde{\mu}_{j,k}(t), z^* - \delta) \leq \tilde{\tau}_{j,k}(t) \operatorname{kl}\left(\tilde{\mu}_{j,k}(t), \tilde{b}_{j,k}(t)\right) \leq \log T + 4\log(\log T). \tag{110}$$

Here the first inequality is from Equation (107). The second inequality stems from Equation (108) and the fact that $\forall 0 < x < 1$, $y \mapsto \operatorname{kl}(x, y)$ is an increasing function when $x < y < 1$ (guaranteed by Equation (109)). The last inequality is according to the definition of $\tilde{b}_{j,k}(t)$. By the definition of s_0, we can get that

$$\operatorname{kl}(\tilde{\mu}_{j,k}(t), z^* - \delta) \leq \operatorname{kl}(\mu_k + \delta, z^* - \delta). \tag{111}$$

As a result, $\tilde{\mu}_{j,k}(t) \geq \mu_k + \delta$ by the fact that $x < y < 1$, $x \mapsto \operatorname{kl}(x, y)$ is a decreasing function when $0 < x < 1$. As a result, $t \in \mathcal{G}_{j,k,1}$, which leads to a contradiction.

(2) We show that $\mathbb{E}[|\mathcal{G}_{j,k,1}|] \leq 5 + 2\delta^{-2}$.

Let $H = \{t \in \mathcal{G}_{j,k} : |\hat{\mu}_{j,k}(t) - \mu_k| \geq \delta\}$. According to Lemma E.1 with $C_t = c = 1$, we have that $\mathbb{E}[H] \leq 4 + 2\delta^{-2}$. Since $k \notin \mathcal{M}^* = \tilde{\mathcal{M}}_j(t)$, arm k can be chosen at most one time in the corresponding block *w.r.t.* t. As a result, $\mathbb{E}[|\mathcal{G}_{j,k,1}|] \leq \mathbb{E}[|H|] \leq 5 + 2\delta^{-2}$.

(3) We show that $\mathbb{E}[|\mathcal{G}_{j,k,2}|] \leq (\log T + 4\log(\log T))/\operatorname{kl}(\mu_k + \delta, z^* - \delta)$.

By the construction of $w(t)$, we have $\mathbb{E}[|\mathcal{G}_{j,k,2}|] \leq s_0$.

Finally, we can prove that

$$\mathbb{E}[|\mathcal{G}_{j,k}|] \leq \mathbb{E}[|\mathcal{G}_{j,k,1}|] + \mathbb{E}[\mathcal{G}_{j,k,2}] \leq \frac{\log T + 4\log(\log T)}{\operatorname{kl}(\mu_k + \delta, z^* - \delta)} + 5 + 2\delta^{-2}. \tag{112}$$

□

C.4. Proofs in Section A

C.4.1. Proof of Proposition A.1

Proof. Let A denote the event that Assumption 3.1 holds and B denote the event that for all $i, j \in [K]$ and $i \neq j$, μ_i/μ_j is an irrational number. Let B_{ij} denote the event that μ_i/μ_j is an irrational number. It is obvious that

$$\bigcap_{1 \leq i < j \leq K} B_{ij} = B \subseteq A. \tag{113}$$

As a result,

$$\mathbb{P}(A) \geq \mathbb{P}(B) = 1 - \mathbb{P}(\bar{B}) = 1 - \mathbb{P}\left(\bigcup_{1 \leq i < j \leq K} \bar{B}_{i,j}\right) \geq 1 - \sum_{1 \leq i < j \leq K} \mathbb{P}(\bar{B}_{ij}). \tag{114}$$

In addition,

$$\mathbb{P}(\bar{B}_{ij}) = \mathbb{E}[\mathbb{P}(\mu_j/\mu \text{ is rational})|\mu_i = \mu]. \tag{115}$$

For a fixed $\mu \in (0, 1]$, since rational number is countable, let $a_1, a_2, a_3, \dots,$ be the rational numbers in $(0, 1/\mu]$. Fix an $\epsilon > 0$ and define the following intervals $[l_1, u_1], [l_2, u_2], \dots$ with

$$l_t = a_t - \epsilon/2^t \quad \text{and} \quad r_t = a_t + \epsilon/2^t. \tag{116}$$

As a result,

$$\mathbb{P}(\mu_j/\mu \text{ is rational}) = \mathbb{P}(\mu_j/\mu \in \{a_1, a_2, \dots\}) \leq \mathbb{P}\left(\mu_j/\mu \in \bigcup_{t \geq 1} [l_t, u_t]\right) \leq \sum_{t \geq 1} \mathbb{P}\left(\mu_j/\mu \in [l_t, u_t]\right). \tag{117}$$

Since the density ratio of distribution D is bounded by M, we have

$$\sum_{t \geq 1} \mathbb{P}\left(\mu_j/\mu \in [l_t, u_t]\right) \leq \sum_{t \geq 1} \mathbb{P}\left(\mu_j \in [l_t\mu, u_t\mu]\right) \leq M\mu \sum_{t \geq 1} \epsilon/2^{t-1} \leq 2M\epsilon. \tag{118}$$

By letting $\epsilon \to 0$, we have $\mathbb{P}(\mu_j/\mu \text{ is rational}) = 0$. Now combining Equations (114) and (115), we have $\mathbb{P}(A) \geq 1$. Now the claim follows. □

C.4.2. Proof of Proposition A.2

Proof. Let $\boldsymbol{s}$ be a Nash equilibrium in Theorem 3.1. Suppose there exists a beneficial deviation for players $B \in [N]$ and the corresponding strategies $\boldsymbol{s}'_B$.

We first prove that $\boldsymbol{s}'_B \subseteq \mathcal{M}^*$. Suppose there exists a player $j \in B$ and $s'_j \notin \mathcal{M}^*$. Then $\mathcal{U}_j^{\text{single}}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \mu_{s'_j} < z^*$ by Theorem 3.1. However, since the original profile $\boldsymbol{s}$ is a Nash equilibrium, we have $\mathcal{U}_j^{\text{single}}(\boldsymbol{s}) \geq z^*$. This leads to the contradiction with the definition of beneficial deviation. As a result, $\boldsymbol{s}'_B \subseteq \mathcal{M}^*$.

We now show that $\boldsymbol{m}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \boldsymbol{m}(\boldsymbol{s})$. Suppose $\boldsymbol{m}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) \neq \boldsymbol{m}(\boldsymbol{s})$, there must exist $k \in \mathcal{M}^*$ such that $m_k(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) > m_k(\boldsymbol{s}) = m_k^*$. As a result, there must exist a player $j \in B$ such that $s'_j = k$ and $\mathcal{U}_j^{\text{single}}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \mu_k/m_k(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) < \mu_k/m_k^* < z^*$. However, since the original profile $\boldsymbol{s}$ is a Nash equilibrium, we have $\mathcal{U}_j^{\text{single}}(\boldsymbol{s}) \geq z^*$. This leads to the contradiction with the definition of beneficial deviation. As a result, $\boldsymbol{m}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \boldsymbol{m}(\boldsymbol{s})$.

Finally, we show that such beneficial deviation does not exist. Base on the previous two claims, we have $\boldsymbol{m}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \boldsymbol{m}(\boldsymbol{s})$. As a result, the rewards earned by the players in B after deviation are a permutation of the original rewards earned by them. Therefore,

$$\sum_{j \in B} \mathcal{U}_j^{\text{single}}(\boldsymbol{s}'_B, \boldsymbol{s}_{-B}) = \sum_{j \in B} \mathcal{U}_j^{\text{single}}(\boldsymbol{s}), \tag{119}$$

which contradicts to the definition of the beneficial deviation which requires that at least one inequality holds strictly for players in B. Now the claim follows. □

C.4.3. Proof of Proposition A.3

Proof. Since $\boldsymbol{\sigma}$ is a symmetric mixed Nash equilibrium, we suppose $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = \cdots = \boldsymbol{\sigma}_N = \boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ where p_k represents the probability for each player to pull arm k. Let $\mathcal{P} = \{k : p_k > 0\}$ be the set of all arms that can be chosen according to $\boldsymbol{p}$. It is obvious that, if $\mathcal{P} \subseteq \mathcal{M}^*$, we must have $W(\boldsymbol{\sigma}) \leq W^{\text{PNE}}$. Now we consider the $\mathcal{P} \backslash \mathcal{M}^* \neq \emptyset$ case.

Since $\boldsymbol{\sigma}$ is a Mixed Nash equilibrium, we have

$$\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(\boldsymbol{s})\right] = \sum_{k\in\mathcal{P}} p_k \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right] \leq \sum_{k\in\mathcal{P}} p_k \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(\boldsymbol{s})\right] = \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(\boldsymbol{s})\right]. \tag{120}$$

This indicates that all the inequalities should be equations in the second step and we have that there exists a constant $C > 0$, such that

$$\forall k \in \mathcal{P}, \quad \mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right] = C. \tag{121}$$

Now we analyze the $\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right]$ term for any $k \in \mathcal{P}$. Let $\boldsymbol{m}_{-j}(\boldsymbol{s}) = \{m_{-j,k}(\boldsymbol{s})\}_{k=1}^K$, where $m_{-j,k}(\boldsymbol{s})$ denotes the number of players except player j that choose arm k. Then $\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right]$ can be expanded as follows:

$$\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right] = \sum_{t=0}^{N-1} \frac{\mu_k}{t+1} \mathbb{P}[m_{-j,k}(\boldsymbol{s}) = t] = \mu_k \sum_{t=0}^{N-1} \frac{1}{t+1} \binom{N-1}{t} p_k^t (1-p_k)^{N-1-t}. \tag{122}$$

To calculate the above equation, we define the following generating function,

$$F(x) = (p_k x + (1-p_k))^{N-1} = \sum_{t=0}^{N-1} a_t x^t, \tag{123}$$

where $a_t = \binom{N-1}{t} p_k^t (1-p_k)^{N-1-t}$. As a result, Equation (122) can be calculated as follows.

$$\mathbb{E}_{\boldsymbol{s}\in\boldsymbol{\sigma}}\left[\mathcal{U}_j^{\text{single}}(k, \boldsymbol{s}_{-j})\right] = \mu_k \int_0^1 F(x)\mathrm{d}x = \mu_k \cdot \left.\frac{(p_k x + (1-p_k))^N}{p_k N}\right|_{x=0}^1 = \frac{\mu_k\left(1-(1-p_k)^N\right)}{p_k N} = C. \tag{124}$$

Hence, $\mu_k\left(1-(1-p_k)^N\right) = C p_k N$ and

$$W(\boldsymbol{\sigma}) = \sum_{k=1}^K \mu_k \mathbb{P}_{\boldsymbol{s}\sim\boldsymbol{\sigma}}[m_k(\boldsymbol{s}) > 0] = \sum_{k=1}^K \mu_k\left(1-(1-\mu_k)^N\right) = \sum_{k\in\mathcal{P}} C p_k N = CN. \tag{125}$$

In addition, Equation (124) indicates that

$$\frac{1-(1-p_k)^N}{p_k} = \frac{CN}{\mu_k}. \tag{126}$$

It is easy to check that function $(1-(1-p_k)^N)/p_k$ is non-increasing and is bounded in $[1, N]$ when $0 \leq p_k \leq 1$. As a result,

$$\forall k \in \mathcal{P}, \quad \frac{CN}{\mu_k} \leq N. \tag{127}$$

Hence we have $CN \leq \mu_k N$ for any $k \in \mathcal{P}$. Combining Equation (126), we have

$$W(\boldsymbol{\sigma}) \leq N\mu_k, \quad \text{for any } k \in \mathcal{P}. \tag{128}$$

Since we consider $\mathcal{P}\backslash\mathcal{M}^*$ case, there exists an arm $k \in \mathcal{P}\backslash\mathcal{M}^*$. Since it is not in the set $\mathcal{M}^*$, according to Theorem 3.1, we have $\mu_k < z^*$. As a result,

$$W(\boldsymbol{\sigma}) \leq N\mu_k \leq Nz^* = \sum_{k\in\mathcal{M}^*} m_k^* z^* \leq \sum_{k\in\mathcal{M}^*} \mu_k = W^{\text{PNE}}. \tag{129}$$

Now the claim follows. □

C.4.4. Proof of Proposition A.5

Proof. Consider the set $\mathcal{J}$ defined in Equation (40), according to Lemma C.3 and Lemma C.4, we have

$$\mathbb{E}[|\mathcal{J}|] \leq \sum_{j=1}^{N} \mathbb{E}[|\mathcal{A}_j \cup \mathcal{B}_j|] \leq \sum_{j=1}^{N} \mathbb{E}[|\mathcal{B}_j \cup \mathcal{C}_j \cup \mathcal{D}_j|] \leq 8N^3K(12K + \delta^{-2}). \tag{130}$$

Since $\mathbb{E}[|\mathcal{J}|] = O(1)$, the expected number of blocks that contain rounds $t \in \mathcal{J}$ is also in $O(1)$. As a result, the differences between $\mathbb{E}[\mathrm{Reg}_j(T)]$ (defined in Equation (4)) and $\mathbb{E}[\mathrm{Reg}'_j(T)]$ in these blocks are also $O(1)$.

Now we analyze the differences between the two regrets for the blocks of which all rounds are not in $\mathcal{J}$. Suppose a block with rounds $vN, vN+1, vN+2, \ldots v \cdot (N+1)$ for an integer v and $(vN+s) \notin \mathcal{J}$ for any $1 \leq s \leq N$. According to Lemma C.2, each player j at these rounds calculates the correct $\tilde{l}_j(t) = l^*$. In addition, at most one player deviates according to Algorithm 1 to explore sub-optimal arms. As a result, the optimal arm for player j at each round in the block is exactly the arm that he plans to pull without the exploration in Lines 16-17 in Algorithm 1. Formally, for each $t \in \{vN, vN+1, \ldots, v \cdot (N+1)\}$,

$$\max_{k \in [K]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} = \frac{\mu_k}{m_k^*}. \tag{131}$$

Since agent j calculates the correct Nash equilibrium and $\tilde{l}_j(t) = l^*$, each arm $k \in \mathcal{M}^*$ is chosen for m_k^* times in $\tilde{l}_j(t)$. As a result,

$$\sum_{t=vN+1}^{v\cdot(N+1)} \max_{k \in [K]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} = \sum_{k \in \mathcal{M}^*} \frac{\mu_k}{m_k^*} \cdot m_k^* = \sum_{k \in \mathcal{M}^*} \mu_k = N \cdot r^* = \sum_{t=vN+1}^{v\cdot(N+1)} r^*. \tag{132}$$

As a result, the differences between the two regrets $\mathbb{E}[\mathrm{Reg}_j(T)]$ and $\mathbb{E}[\mathrm{Reg}'_j(T)]$ are 0 in these blocks.

Combining these results, we know that $\mathbb{E}[\mathrm{Reg}_j(T)]$ and $\mathbb{E}[\mathrm{Reg}'_j(T)]$ have differences at most $O(1)$. According to Theorem 3.3, we have

$$\mathbb{E}[\mathrm{Reg}'_j(T)] \leq \mathbb{E}[\mathrm{Reg}_j(T)] + O(1) \leq \sum_{k \notin \mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + O(1). \tag{133}$$

Now the claim follows. □

C.4.5. Proof of Proposition A.6

Proof. By the no regret condition, we have $\sum_{j=1}^{N} \mathbb{E}[\mathrm{Reg}_j(T)] \leq o(T^\alpha)$. We note that each non-equilibrium round contributes at least δ_0 regret since there must exist at least one player that can gain δ_0 increase in reward by deviation. As a result,

$$\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_j(T)] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{N}\left(\max_{k \in [K]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right)\right] \\
&\geq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{N}\left(\max_{k \in [K]} \frac{\mu_k}{M_k(t) + \mathbb{I}[\pi_j(t) \neq k]} - R_j(t)\right) \middle| \exists k \in [K], M_k^*(t) \neq m_k^*\right] \mathbb{P}\left[\exists k \in [K], M_k^*(t) \neq m_k^*\right] \\
&\geq \delta_0 \,\mathrm{NonEqu}(T).
\end{aligned} \tag{134}$$

As a result, $\mathrm{NonEqu}(T) \leq o(T^\alpha)$. Therefore, for all $k \in \mathcal{M}^*$,

$$\mathbb{E}\left[\sum_{j=1}^{N} \tau_{j,k}(T)\right] \geq m_k^* \left(T - \mathbb{E}[\mathrm{NonEqu}(T)]\right). \tag{135}$$

Algorithm 2 SMAA (without knowledge of N and rank)

$\hat{N}, j \leftarrow$ Musical Chairs Approach ((Rosenski et al., 2016), Algorithm 3)
$T_1 \leftarrow \lceil 50K^2 \log(4T) \rceil + \left\lceil \hat{N} \log(2T) \right\rceil$
Call Algorithm 1 for t from $T_1 + 1$ to T.

Algorithm 3 Musical Chairs (Rosenski et al., 2016)

1: $T_0 \leftarrow \lceil 50 \log(4T) K^2 \rceil$
2: $C \leftarrow 0$
3: **for** $t \leftarrow 1$ to T_0 **do**
4: Sample arm k uniformly in $[K]$
5: Observe the collision information $\eta(t)$
6: **if** $\eta(t) = 1$ **then**
7: $C \leftarrow C + 1$
8: **end if**
9: **end for**
10: $\hat{N} \leftarrow \min \left\{ \text{round} \left(\frac{\log((T_0 - C)/T_0)}{\log(1 - 1/K)} + 1 \right), K \right\}$ and $\hat{N} \leftarrow K$ if $C = T_0$
11: $j \leftarrow -1$
12: $T_1 = T_0 + \left\lceil \hat{N} \log(2T) \right\rceil$
13: **for** $t \leftarrow T_0 + 1$ to T_1 **do**
14: **if** $j = -1$ **then**
15: Sample i uniformly in $[\hat{N}]$ and pull arm i
16: Observe the collision information $\eta(t)$
17: **if** $\eta(t) = 0$ **then**
18: $j \leftarrow i$
19: **end if**
20: **else**
21: Pull arm j
22: **end if**
23: **end for**
24: **Output:** $\hat{N}$, j

Therefore, $m_k^* T - \sum_{j=1}^{N} \mathbb{E}[\tau_{j,k}(T)] \leq o(T^\alpha)$. Now due to the fairness condition, we have that for all $\alpha > 0$, $j \in [N]$ and $k \in \mathcal{M}^*$,

$$\frac{m_k^*}{N} T - \mathbb{E}[\tau_{j,k}(T)] \leq o(T^\alpha). \tag{136}$$

The claim follows. □

D. More Details about the Setting When N and Rank is unknown

D.1. Algorithm Details

The pseudo-code of the whole method is shown in Algorithm 2 and the pseudo-code of the Musical Chairs approach is shown in Algorithm 3. When $t \in [T_0]$, players randomly choose the arms in $[K]$ and count the number of collisions. The number of players is then estimated according to Line 10 in Algorithm 3. When $t \in \{T_0 + 1, \ldots, T\}$, each player first randomly chooses the arms in $[\hat{N}]$. If no collision occurs, the player will hold on to the arm in the remaining rounds. Otherwise, the player re-sampleso an arm uniformly in $[\hat{N}]$ and the progress goes on. The output rank of the player is the index of the arm he holds on.

D.2. Properties

We first show that the strategic player will not affect other players that follow the Musical Chairs approach in Proposition D.1.

Proposition D.1. *If $N - 1$ players follow the Musical Chairs approach, then with probability at least $1 - N/T$, after*

$T_1 = \lceil 50K^2 \log(4T) \rceil + \lceil N \log(2T) \rceil$ rounds, regardless of the policy of the strategic player, each player could estimate the number of players N accurately and get a different rank $j \in [N]$ with other players except the strategic player.

Similar to Algorithm 1, we could show the properties of Algorithm 2 by the following theorem (formal version of Corollary 3.7).

Theorem D.2. *Under Assumption 3.1, suppose $0 < \delta < \delta_0/2$. Let $\mu_{\max} = \max_{k\in[K]} \mu_k$. Then Algorithm 2 satisfies the following properties.*

1. *When all players follow Algorithm 2, the expected regret for each player j is upper-bounded by the following equation.*

$$\mathbb{E}[\mathrm{Reg}_j(T)] \leq \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\mu_{\max}\log(4T) + 10N^3K(13K + \delta^{-2}) + N\mu_{\max}. \tag{137}$$

2. *When all players follow Algorithm 2, the expected number of rounds that are not a Nash equilibrium demonstrated in Theorem 3.1 is upper bounded by the following equation.*

$$\mathbb{E}[\mathrm{NonEqu}(T)] \leq N \sum_{k\notin\mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\log(4T) + 10N^3K(14K + \delta^{-2}). \tag{138}$$

3. *The policy profile where all players follow Algorithm 2 is an ϵ-Nash equilibrium w.r.t. to the game specified by $(\mathcal{S}_{policy}, \{\mathcal{U}_j^{policy}\}_{j=1}^N)$ and is $(\beta, \epsilon + \beta\gamma)$-stable with*

$$\begin{aligned}
\beta &= \frac{\delta_0}{z^*}, \\
\epsilon &= \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\mu_{\max}\log(4T) + 10N^3K(13K + \delta^{-2}) + N\mu_{\max}, \\
\gamma &= \sum_{k\notin\mathcal{M}^*} \frac{\log T + \log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\mu_{\max}\log(4T) + 10N^3K(13K + \delta^{-2}) + N\mu_{\max}.
\end{aligned} \tag{139}$$

Here z^* is provided in Equation (8).

D.3. Proofs

D.3.1. Proof of Proposition D.1

Proof. At each round $t \in [T_0]$, suppose the probability of the strategic behavior to pull arm k is $q_k(t)$. As a result, the probability of any other player j to observe a collision is

$$\begin{aligned}
P(\eta(t) = 1) &= \sum_{k=1}^{K} P(\pi_j(t) = k) P(\eta(t) = 1 | \pi_j(t) = k) \\
&= \sum_{k=1}^{K} \frac{1}{K} \cdot \left(1 - \left(1 - \frac{1}{K}\right)^{N-2} (1 - q_k(t))\right) \\
&= 1 - \left(1 - \frac{1}{K}\right)^{N-2} \sum_{k=1}^{K} \frac{1 - q_k(t)}{K} \\
&= 1 - \left(1 - \frac{1}{K}\right)^{N-1}.
\end{aligned} \tag{140}$$

As a result, no matter the behaviors of the strategic player, the probability of other players that follow Algorithm 3 to observe a collision is constant. According to Lemma 3 in (Rosenski et al., 2016), we can get that if $T_0 = \lceil 50\log(4T)K^2 \rceil$, then with probability at least $1 - 1/(2T)$, we have $\hat{N} = N$.

For any $t \in \{T_0+1, \dots, T_1\}$, when $\hat{N} = N$, then the probability of a player j that follows Algorithm 3 to have $\eta_j(t) = 0$ is at least $1/N$ since there must exist at least one arm that is not pulled in $[N]$. As a result, when $\hat{N} = N$, since $T_1 - T_0 = \left\lceil \hat{N}\log(2T) \right\rceil$, the probability for player j to always meet collisions is

$$\mathbb{P}\left[\forall t \in \{T_0+1, \dots, T_1\}, \eta_j(t) = 0\right] \leq \left(1 - \frac{1}{N}\right)^{T_1 - T_0} \leq \exp\left(-\frac{T_1 - T_0}{N}\right) \leq \frac{1}{2T}. \tag{141}$$

As a result, when $T_0 = \left\lceil 50\log(4T)K^2 \right\rceil$ and $T_1 - T_0 = \left\lceil \hat{N}\log(2T) \right\rceil$, the probability of a player that follows Algorithm 3 to estimate N accurately and get a different rank with other players is at least $(1 - 1/(2T))^2 \leq 1 - 1/T$. Now the claim follows by applying the union bound for all players. □

D.3.2. Proof of Theorem D.2

Proof. For the first part, let $\mathcal{E}$ denote the clean event when all players that follow Algorithm 2 estimate N accurately and get a different rank j. Let $\mu_{max} = \max_{k\in[K]} \mu_k$ and $\mathrm{Reg}_j^{\mathrm{ee}}(T)$ represent the regret in the exploration-exploitation phase (Algorithm 1). As a result,

$$\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_j(T)] \\
=&\mathbb{P}[\mathcal{E}]\mathbb{E}\left[\mathrm{Reg}_j(T)|\mathcal{E}\right] + \mathbb{P}[\bar{\mathcal{E}}]\mathbb{E}\left[\mathrm{Reg}_j(T)|\bar{\mathcal{E}}\right] \\
\leq& T_1\mu_{\max} + \mathrm{Reg}_j^{\mathrm{ee}}(T) + \frac{N}{T}\mu_{\max}\cdot T \\
\leq& \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\mu_{\max}\log(4T) + 10N^3K(13K + \delta^{-2}) + N\mu_{\max}.
\end{aligned} \tag{142}$$

For the second part, let $\mathrm{NonEqu}^{\mathrm{ee}}(T)$ represent the rounds of non-equilibrium in the exploration-exploitation phase (Algorithm 1). Then we have

$$\begin{aligned}
&\mathrm{NonEqu}(T) \\
=&\mathbb{P}[\mathcal{E}]\mathbb{E}\left[\mathrm{NonEqu}(T)|\mathcal{E}\right] + \mathbb{P}[\bar{\mathcal{E}}]\mathbb{E}\left[\mathrm{NonEqu}(T)|\bar{\mathcal{E}}\right] \\
\leq& T_1 + \mathrm{NonEqu}^{\mathrm{ee}}(T) + \frac{N}{T}\cdot T \\
\leq& N\sum_{k\notin\mathcal{M}^*} \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\log(4T) + 10N^3K(14K + \delta^{-2}).
\end{aligned} \tag{143}$$

For the third part, according to Proposition D.1, the behaviors of the selfish player in the Musical Chairs phase will not affect the results of other players. In addition, as shown in the proof of Theorem 3.5 in Appendix C.3.4, Algorithm 1 is not affected by selfish players. As a result, we conclude that the policy profile where all players follow Algorithm 2 is an ϵ-Nash equilibrium with

$$\epsilon = \sum_{k\notin\mathcal{M}^*} \frac{(z^* - \mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \delta, z^* - \delta)} + (50K^2 + N)\mu_{\max}\log(4T) + 10N^3K(13K + \delta^{-2}) + N\mu_{\max}. \tag{144}$$

Let $\mathrm{Rew}_j^{\mathrm{ee}}(T; \boldsymbol{s})$ represents the reward of player j in the exploration-exploitation phase (Algorithm 1 part) under the original profile $\boldsymbol{s}$ where all players follow Algorithm 2. When a selfish player deviate from s to s'. Let $\mathcal{Q}$ be the set of rounds t that player j deviates from the original algorithm in the exploitation-exploration phase and all other players calculate the correct Nash equilibrium (*i.e.*, $\mathcal{Q}_3$ in the proof of Theorem 3.5 in Appendix C.3.4). The reward of player j and a player i

that follows Algorithm 2 is given by the following equations.

$$\begin{aligned}
&\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_i(T; \boldsymbol{s})] \\
=&\mathbb{P}[\mathcal{E}]\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_i(T; \boldsymbol{s})|\mathcal{E}] + \mathbb{P}[\bar{\mathcal{E}}]\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_i(T; \boldsymbol{s})|\bar{\mathcal{E}}] \\
\geq& -T_1\mu_{\max} - \mathbb{E}[\mathrm{Rew}_i^{\mathrm{ee}}(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_i^{\mathrm{ee}}(T; \boldsymbol{s})|\bar{\mathcal{E}}] - \frac{N}{T}\cdot T\mu_{\max} \\
\geq& -10N^3K(13K+\delta^{-2}) - \sum_{k\notin\mathcal{M}^*}\frac{\log T + \log(\log T)}{\mathrm{kl}(\mu_k+\delta, z^*-\delta)} - (50K^2+N)\mu_{\max}\log(4T) - N\mu_{\max} - z^*\mathbb{E}[|\mathcal{Q}|] \\
=& -\gamma - z^*\mathbb{E}[|\mathcal{Q}|].
\end{aligned} \tag{145}$$

And

$$\begin{aligned}
&\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_j(T; \boldsymbol{s})] \\
=&\mathbb{P}[\mathcal{E}]\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_j(T; \boldsymbol{s})|\mathcal{E}] + \mathbb{P}[\bar{\mathcal{E}}]\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_j(T; \boldsymbol{s})|\bar{\mathcal{E}}] \\
\leq& T_1\mu_{\max} + \mathbb{E}[\mathrm{Rew}_i^{\mathrm{ee}}(T; s', \boldsymbol{s}_{-j}) - \mathrm{Rew}_i^{\mathrm{ee}}(T; \boldsymbol{s})|\bar{\mathcal{E}}] + \frac{N}{T}\cdot T\mu_{\max} \\
\leq& 10N^3K(13K+\delta^{-2}) + \sum_{k\notin\mathcal{M}^*}\frac{(z^*-\mu_k)(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k+\delta, z^*-\delta)} - \delta_0\mathbb{E}[|\mathcal{Q}|] + (50K^2+N)\mu_{\max}\log(4T) + N\mu_{\max} \\
=& \epsilon - \delta_0\mathbb{E}[|\mathcal{Q}|].
\end{aligned} \tag{146}$$

As a result, for any $u \in \mathbb{R}^+$, if $\mathbb{E}[\mathrm{Rew}_i(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_i(T; \boldsymbol{s})] \leq -u$, we have $\mathbb{E}[|\mathcal{Q}|] \geq (u-\gamma)/z^*$. Then

$$\mathbb{E}[\mathrm{Rew}_j(T; s', \boldsymbol{s}_{-j})] - \mathbb{E}[\mathrm{Rew}_j(T; \boldsymbol{s})] \leq \epsilon - \delta_0\frac{u-\gamma}{z^*} = \epsilon + \frac{\delta_0\gamma}{z^*} - \frac{\delta_0}{z^*}u. \tag{147}$$

Now the claim follows. □

E. Auxiliary Lemmas

Lemma E.1 (Lemma 5 in (Combes et al., 2015a)). *Let $\mathcal{F}_t$ be the σ-algebra generated by $\{X_k(t), \forall k \in [K]\}_{s=1}^t$. Let $j \in [N]$, $k \in [K]$ and $c > 0$. Let H be a random set of rounds such that for all t, $\mathbb{I}[t \in H] \in \mathcal{F}_{t-1}$. Assume that there exists $\{C_t\}_{t\in\mathbb{N}}$, a sequence of independent binary random variables such that for any $t \geq 1$, C_t is $\mathcal{F}_t$-measurable and $\mathbb{P}[C_t = 1|t \in H] \geq c$. Further assume for any $t \in H$, k is selected ($\pi_j(t) = k$) if $C_t = 1$. Then*

$$\sum_{t\geq 1}\mathbb{P}[t \in H, |\hat{\mu}_{j,k}(t) - \mu_k| \geq \delta] \leq 2c^{-1}\left(2c^{-1} + \delta^{-2}\right). \tag{148}$$

Lemma E.2 (Theorem 10 in (Garivier & Cappé, 2011)). *Let $\{Z_t\}_{t\geq 1}$ be a sequence of independent random variables bounded in $[0,1]$ with common expectation $\mu = \mathbb{E}[Z_t]$. Let $\mathcal{F}$ be an increasing sequence of σ-fields such that for each t, $\sigma(Z_1, Z_2, \ldots, Z_t) \subseteq \mathcal{F}_t$ and for $s > t$, Z_s is independent from $\mathcal{F}_t$. Consider a previsible sequence $\{\epsilon_t\}_{t\geq 1}$ of Bernoulli variables (for all $t > 0$, ϵ_t is $\mathcal{F}_{t-1}$-measurable). Let $\delta > 0$ and for every $u \in \{1, \ldots, t\}$ let*

$$\begin{aligned}
S(u) = \sum_{s=1}^u \epsilon_s X_s, \quad N(u) = \sum_{s=1}^u \epsilon_s, \quad \hat{\mu}(u) = \frac{S(u)}{N(u)}, \\
b(t) = \sup\left\{q \geq \hat{\mu}(t) : N(t)\,\mathrm{kl}(\hat{\mu}(t), q) \leq \delta\right\}.
\end{aligned} \tag{149}$$

Then

$$\mathbb{P}[b(t) < \mu] \leq \lceil \delta\log t\rceil \exp(1-\delta). \tag{150}$$