PPA-Game: Characterizing and Learning Competitive Dynamics Among Online Content Creators

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

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]School of Economics and Management, Tsinghua University, Beijing, China

Abstract

We introduce the Proportional Payoff Allocation Game (PPA-Game) to model how agents, akin to content creators on platforms like YouTube and TikTok, compete for divisible resources and consumers' attention. Payoffs are allocated to agents based on heterogeneous weights, reflecting the diversity in content quality among creators. Our analysis reveals that although a pure Nash equilibrium (PNE) is not guaranteed in every scenario, it is commonly observed, with its absence being rare in our simulations. Beyond analyzing static payoffs, we further discuss the agents' online learning about resource payoffs by integrating a multi-player multi-armed bandit framework. We propose an online algorithm facilitating each agent's maximization of cumulative payoffs over T rounds. Theoretically, we establish that the regret of any agent is bounded by $O(\log^{1+\eta} T)$ for any $\eta > 0$. Empirical results further validate the effectiveness of our approach.

1 Introduction

Online recommender systems, exemplified by YouTube and TikTok, are now central to our daily digital experiences. From an economic viewpoint, these systems represent intricate ecosystems composed of multiple stakeholders: consumers, content creators, and the platform itself [Boutilier et al., 2023, Ben-Porat et al., 2020]. Although a significant portion of current research delves into consumer preferences and individualized recommendations, the role and behavior of content creators are also crucial [Zhan et al., 2021, Boutilier et al., 2023]. They not only shape the pool of recommended content but also influence the system's stability, fairness, and long-term utility [Ben-Porat and Tennenholtz, 2018, Patro et al., 2020, Boutilier et al., 2023, Mladenov et al., 2020]. A major challenge when capturing the dynamics of content creators is characterizing the competitive exposure amongst them, given the finite demand for different content. Consequently, understanding the competitive dynamics among content creators becomes paramount.

Game theory offers a useful framework to characterize competitive behaviors among multiple players [Nisan et al., 2007]. While numerous studies have employed game-theoretical analysis on content creators within online platforms, many [Boutilier et al., 2023, Ben-Porat and Tennenholtz, 2018, Hron et al., 2022, Jagadeesan et al., 2022, Ben-Porat et al., 2019a] primarily explore Nash equilibrium properties, largely overlooking the dynamics of content creators and their convergence to the equilibrium. Some research attempts to bridge this gap, considering different allocation models for creators producing analogous content. Specifically, Ben-Porat et al. [2019b, 2020], Liu et al. [2020, 2021] posit that only top-tier content creators receive exposure, and Xu et al. [2023] proposes an averaged allocation rule. While these dynamics yield persuasive findings, they exhibit limitations stemming from the disconnect between their foundational principles and real-world contexts. For instance, top-winner dynamics may conflict with diversity principles in online recommendation platforms, while an averaged allocation rule could overlook the nuanced specializations that creators bring to diverse content domains.

*Email: xrz199721@gmail.com
†Corresponding author. Email: cuip@tsinghua.edu.cn

To bridge the gap between theoretical models and the complex dynamics observed among online content creators, we introduce the Proportional Payoff Allocation Game (PPA-Game) as a novel model to capture the competitive dynamics of online content creators. Adapting from [Ben-Porat et al., 2019b, 2020, Prasad et al., 2023], we assume a game with N players (content creators) and K resources (content topics). In this context, players need to decide which resources to select. Each resource has an associated total payoff (the exposure for each content topic), and individual players have specific weights (quality metrics) assigned to each resource. When several players choose the same resource, its payoff is distributed proportionally according to player weights—a mechanism resonating with other game-theoretical frameworks [Caragiannis and Voudouris, 2016, Hoefer and Jiamjitrak, 2017, Anbarci et al., 2023].

We first examine the existence of pure Nash equilibria (PNE) in our game model, uncovering that PNE may not always exist in certain scenarios. While this result could initially appear frustrating, our synthetic experiments suggest that instances lacking PNE are infrequent. Further investigation has identified several conditions under which the existence of PNE is typically ensured. These conditions, which have wide-ranging applicability in practical scenarios, ensure the presence of PNE if any of the following criteria are met: (1) The resources' payoffs follow a long-tailed distribution, wherein the payoff ratio between any two resources either surpasses a specific constant c_0 or falls below its reciprocal $1/c_0$. (2) Players have the same weights across different resources. (3) Players have the same weights within each resource. (4) All resources have equal payoffs.

Leveraging the PPA-Game, we introduce a Multi-player Multi-Armed Bandit (MPMAB) framework [Rosenski et al., 2016, Besson and Kaufmann, 2018, Boursier and Perchet, 2019, 2020, Shi et al., 2020, 2021, Huang et al., 2022, Lugosi and Mehrabian, 2022] to simulate environments where content creators lack prior information about content demand and their quality in various topics. In this context, the 'players' represent content creators, while 'arms' denote content topics. This model entails players making simultaneous decisions over multiple rounds, with payoffs determined by the PPA-Game mechanism. Motivated by [Xu et al., 2023, Boursier and Perchet, 2020], we assume that selfish players, motivated by their own interests, aim to maximize their own cumulative payoffs over T rounds. Within this MPMAB structure, inspired by [Bistritz and Leshem, 2018, 2021, Pradelski and Young, 2012], we delineate a novel online learning strategy for players and prove that each player's regret is bounded by $O(\log^{1+\eta} T)$ for any $\eta > 0$. Synthetic experiments validate the effectiveness of our approach.

2 Related Works

Game-theoretical Analysis for Content Creators. While much of the research in this area has focused on consumer preferences and recommendations [Boutilier et al., 2023, Zhan et al., 2021], some studies have examined the competition among content creators using game-theoretical models [Boutilier et al., 2023, Ben-Porat and Tennenholtz, 2018, Hron et al., 2022, Jagadeesan et al., 2022, Ben-Porat et al., 2019a,b, 2020, Xu et al., 2023, Liu et al., 2020, 2021]. A subset of these studies attempts to capture the dynamics among content creators. However, they often oversimplify exposure allocation, either favoring top-quality content creators [Ben-Porat et al., 2019b, 2020, Liu et al., 2020, 2021] or assuming equal exposure distribution [Xu et al., 2023]. In contrast to these approaches, we assume that creators bring distinct expertise and excel in producing varied content types. Furthermore, different from Yao et al. [2023]'s study on coarse correlated equilibria with no-regret dynamics, our work distinctively focuses on pure Nash equilibria.

Proportional Allocation. Proportional allocation is a well-explored mechanism in game theory, with various applications from resource distribution [Kelly, 1997, Caragiannis and Voudouris, 2016, Syrgkanis and Tardos, 2013, Johari, 2007] to Blotto games [Anbarci et al., 2023, Klumpp et al., 2019, Arad and Rubinstein, 2012, Friedman, 1958]. These studies often focus on the strategic allocation of a single divisible resource or budget distribution across multiple contests. In contrast, our research diverges significantly by considering multiple divisible resources, with players needing to make strategic choices among these resources. This approach differs notably from the proportional coalition formation game explored by Hoefer and Jiamjitrak [2017], where payoffs within coalitions are divided based on player weights. Additionally, while the weighted congestion game [Bhawalkar et al., 2014] assumes that players have uniform weights across resources, our model accounts for players having different weights on various resources, adding a layer of complexity and

realism to our game-theoretical analysis.

Multi-player Multi-Armed Bandits. The Multi-Armed Bandit problem and its multi-player variant (MPMAB) have been extensively explored [Lai et al., 1985, Bubeck et al., 2012, Slivkins et al., 2019, Boursier and Perchet, 2022]. In MPMAB, a common assumption is the absence of payoffs during player collisions [Boursier and Perchet, 2022]. Various studies propose different payoff allocation mechanisms for collision scenarios, ranging from awarding the entire payoff to the highest bidder [Liu et al., 2020, 2021, Jagadeesan et al., 2021, Basu et al., 2021] to implementing thresholds or capacities for each arm [Bande and Veeravalli, 2019, Youssef et al., 2021, Bande et al., 2021, Magesh and Veeravalli, 2021, Wang et al., 2022a,b]. Unlike the models that either maximize total welfare or offer reduced payoffs during collisions [Shi and Shen, 2021, Pacchiano et al., 2021, Boyarski et al., 2021], our study focuses on individual payoff maximization, highlighting player selfishness.

3 Proportional Payoff Allocation Game

3.1 Background on Game Theory

Nash Equilibrium. The concepts of ϵ-Nash equilibrium and pure Nash equilibrium (PNE) have been well-established in the game theory literature [Nisan et al., 2007]. To set the context for our study, consider a game with N players. Let $\mathcal{S}$ represent the set of potential strategies for these players. Each player j adopts a strategy $\pi_j \in \mathcal{S}$. The collective strategy of all players is represented as $\boldsymbol{\pi} = \{\pi_j\}_{j=1}^N$, while $(\pi', \boldsymbol{\pi}_{-j})$ denotes a scenario where player j deviates from his original strategy π_j to another strategy $\pi' \in \mathcal{S}$. We denote the payoff of player j for a given strategy profile $\boldsymbol{\pi}$ as $\mathcal{U}_j(\boldsymbol{\pi})$. With these notations, the ϵ-Nash equilibrium is defined as:

Definition 3.1 (ϵ-Nash Equilibrium and Pure Nash Equilibrium (PNE))**.** A strategy profile $\boldsymbol{\pi} \in \mathcal{S}^N$ for a game defined by $\mathcal{G} = \left(\mathcal{S}, \{\mathcal{U}_j\}_{j=1}^N\right)$ is termed an ϵ-Nash equilibrium if, for every strategy $\pi' \in \mathcal{S}$ and for each player $j \in [N]$, $\mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) \leq \mathcal{U}_j(\boldsymbol{\pi}) + \epsilon$. If $\boldsymbol{\pi}$ achieves a 0-Nash equilibrium, it's termed a pure Nash equilibrium (PNE).

Total Welfare. The total welfare, represented by $W_{\mathcal{G}}(\boldsymbol{\pi})$, calculates the aggregate payoff derived from a strategy profile $\boldsymbol{\pi}$. In the context of our game, it's the summation of the utilities obtained by all participating players, expressed mathematically as: $W_{\mathcal{G}}(\boldsymbol{\pi}) = \sum_{j=1}^N \mathcal{U}_j(\boldsymbol{\pi})$.

Improvement Step and Improvement Path. We leverage the notions of improvement steps and paths from prior works [Ben-Porat et al., 2019b, 2020]. Starting with any strategy profile $\boldsymbol{\pi} \in \mathcal{S}^N$, an *improvement step* is a transition to a new strategy profile $(\pi'_j, \boldsymbol{\pi}_{-j})$, where player j unilaterally deviates to gain a higher payoff, such that $\mathcal{U}_j(\pi'_j, \boldsymbol{\pi}_{-j}) > \mathcal{U}_j(\boldsymbol{\pi})$. Furthermore, an *improvement path* $\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \dots)$ is a sequence formed by chaining such improvement steps. Given our focus on a finite game with bounded player counts and strategy space (as shown in Definition 3.2), any unending improvement path will inevitably contain cycles, implying the existence of states $t_1 < t_2$ where $\boldsymbol{\pi}(t_1) = \boldsymbol{\pi}(t_2)$. Crucially, if all potential improvement paths in a game are finite, every better-response dynamics will converge to a PNE [Nisan et al., 2007].

3.2 Game Formulation

Consider a scenario with K distinct resources labeled by indices $[K] = \{1, 2, \dots, K\}$ and N players, where $N, K \geq 2$. Each resource, denoted by k, possesses a total payoff μ_k in the range $0 < \mu_k \leq 1$. Simultaneously, each player, represented as j, has a weight ranging between $0 \leq w_{j,k} \leq 1$ for resource k. Every player can choose from a strategy set $[K]$, resulting in an overall strategy profile $\boldsymbol{\pi} \in [K]^N$. Crucially, if multiple players opt for the same resource, the total payoff from that resource is distributed proportionally based on the individual player weights. This can be formalized as:

Definition 3.2 (Proportional Payoff Allocation Game (PPA-Game))**.** Given K resources with associated payoffs $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^K$ $(0 < \mu_k \leq 1)$ and N players characterized by weights $\boldsymbol{W} = \{w_{j,k} : j \in$

$[N], k \in [K]\}$ $(0 \leq w_{j,k} \leq 1)$, the Proportional Payoff Allocation Game (PPA-Game) is formulated as $\mathcal{G} = \left(\mathcal{S}, \{\mathcal{U}_j(\boldsymbol{\pi})\}_{j=1}^{N}\right)$, where $\mathcal{S} = [K]$ and

$$\mathcal{U}_j(\boldsymbol{\pi}) = \begin{cases} \frac{\mu_k w_{j,k}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'}=k] w_{j',k}} & \text{if } \sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = k] w_{j',k} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here $k = \pi_j$ and $\mathbb{I}[\cdot]$ is the indicator function.

It's worth noting that any resource k without players having positive weights will remain unchosen. Consequently, such a resource can be disregarded. Based on this observation, we can assert that for each $k \in [K]$, there exists a $j \in [N]$ such that $w_{j,k} > 0$, and the maximum weight across players is 1, *i.e.*, $\max_{j \in [N]} w_{j,k} = 1$ for every $k \in [K]$. We then demonstrate that how the PPA-Game can be used to model competitive behaviors among content creators within online recommender systems:

Example 3.1. In modern recommendation platforms like YouTube and TikTok, content creators compete for viewer attention by producing content on diverse topics. In this context, content creators serve as the 'players,' while the various topics they generate act as the 'resources.' The 'payoff' associated with each resource represents the overall attention or exposure a topic receives. Since creators possess varying expertise across different topics, the exposure tends to be proportionally distributed based on their respective influence or proficiency in that topic.

Several models closely relate to our formulation, particularly the studies by [Bhawalkar et al., 2014, Ben-Porat et al., 2019b, 2020, Liu et al., 2020, 2021, Xu et al., 2023]. However, models from [Ben-Porat et al., 2019b, 2020, Liu et al., 2020, 2021] assume that only the player j with the highest weight $w_{j,k}$ for a resource k receives the payoff. Xu et al. [2023] assumed an average payoff allocation to players choosing the resource. The weighted congestion game [Bhawalkar et al., 2014] assumes that players have the same weights on all resources. Hence, we note that all these formulations diverge from realistic scenarios of online recommender systems, as explained above.

3.3 Analysis on Pure Nash Equilibrium (PNE)

In this subsection, we investigate Pure Nash Equilibria (PNEs) due to their role as the most stringent equilibria in game theory [Nisan et al., 2007] and their importance for the stability of platforms [Ben-Porat and Tennenholtz, 2018]. This exploration is consistent with previous research [Ben-Porat et al., 2020, Xu et al., 2023, Hron et al., 2022] that has examined PNEs within recommender systems.

3.3.1 High Probability of PNE Existence

We first acknowledge that PNE may not always exist in certain scenarios (see Appendix A.1 for detailed illustrative example):

Observation 3.1. PNE does not exist in specific scenarios.

While this result might seem frustrating, subsequent synthetic experiments largely demonstrate the typical existence of PNEs in most cases in a large-scale simulation study (see Appendix A.2 for detailed experimental setup):

Observation 3.2. We randomly choose N and K from the range $[10, 100]$ and sample resources' payoffs and players' weights from four distributions supported on $[0, 1]$. Among the $20,000,000$ randomly generated game configurations we tested, PNE was absent in only one instance.

These synthetic experiments allow us to infer that the existence of PNE is a common occurrence in practical scenarios.

3.3.2 Existence of PNE in Specific Scenarios

While judging the existence of PNE in general cases is impossible, we prove that PNE exists in several typical scenarios, and each corresponds to some realistic case:

Theorem 3.1 (Long-tailed Resource Scenario)**.** *Define N_0 and ϵ_0 as follows.*

$$N_0 = \max_{k\in[K]} \sum_{j=1}^{N} \mathbb{I}[w_{j,k} > 0], \qquad \epsilon_0 = \min\left\{w_{j,k} : j \in [N], k \in [K], w_{j,k} > 0\right\}. \tag{2}$$

Then the PNE exists if

$$\forall k \le k', \quad \frac{\mu_k}{\mu_{k'}} > \frac{N_0}{\epsilon_0} \quad \text{or} \quad \frac{\mu'_k}{\mu_k} > \frac{N_0}{\epsilon_0}. \tag{3}$$

Remark 3.1. The proof is provided in Appendix C.1. Theorem 3.1 demonstrates scenarios with a long-tailed distribution of resource payoffs. In particular, Equation (3) specifies that the peak payoff among resources substantially surpass that of other resources. For instance, entertainment-related content on TikTok garners substantially more views than other topics [Statista, 2023].

Theorem 3.2. *The existence of PNE is guaranteed if any of the following conditions are met:*

1. *(Partially Heterogeneous Scenario) Players have the same weights for different resources, i.e., $\forall k, k' \in [K], j \in [N]$, $w_{j,k} = w_{j,k'}$.*
2. *(Homogeneous Player Scenario) Players have the same weights within each resource, i.e., $\forall k \in [K]$ and $j, j' \in [N]$, $w_{j,k} = w_{j',k}$.*
3. *(Homogeneous Resource Scenario) Resources have the same payoffs, i.e., $\forall k, k' \in [K]$, $\mu_k = \mu_{k'}$.*

Remark 3.2. The proof is provided in Appendix C.2. Theorem 3.2 demonstrated that the PNE exists in several degraded cases of the PPA-Game in Definition 3.2. Specifically, Condition (1) corresponds to scenarios where the weights reflect the inherent influence of different players. For instance, in online recommender platforms, certain creators, termed as 'influencers,' have greater impact on all topics owing to their expertise, stature, or authority. We also note that Condition (1) exhibits a special case of the weighted congestion game [Bhawalkar et al., 2014]. However, PNE may not exist in weighted congestion games while we prove the existence in this specific scenario. Condition (2) encompasses cases where players are homogeneous and have the same influence on each topic. This scenario aligns with the one addressed in [Xu et al., 2023]. Condition (3) considers a homogeneous resource scenario wherein all resources have equivalent payoffs. Such scenarios include that users compete for processing resources in modern data centers while each resource has the same processing power [Benblidia et al., 2021].

3.3.3 Non-uniqueness and Inefficiency of PNEs

We further show that PNEs might not be unique. Notably, among multiple PNEs, some can yield a reduced total welfare compared to others:

Example 3.2 (Existence of multiple and inefficient PNEs)**.** Assume the resources' payoffs to be $[0.6, 0.4, 0.2]$. Suppose each player has the same weights on the resources (as per Condition (2) in Theorem 3.1), with weights $[1, 3/8, 1/4]$, respectively. Under such a setup, two distinct PNEs emerge: $\boldsymbol{\pi}^{(1)} = (1, 2, 3)$ and $\boldsymbol{\pi}^{(2)} = (2, 1, 1)$. Notably, since $W_{\mathcal{G}}\left(\boldsymbol{\pi}^{(1)}\right) = 1.2$ exceeds $W_{\mathcal{G}}\left(\boldsymbol{\pi}^{(2)}\right) = 1.0$, the latter, $\boldsymbol{\pi}^{(2)}$, is a less efficient PNE.

Given the existence for such inefficient PNEs, it becomes preferable for players to converge to an optimal PNE, thereby maximizing the overall welfare.

4 Online Learning

In real-world scenarios, the payoffs associated with each resource are not static but fluctuate over time, influenced by varying factors such as temporal demand for different content topics. Furthermore, players

typically lack prior information about both the payoffs of resources and their own respective weights. To capture the learning dynamics of players in the PPA-Game, we introduce a novel framework based on multi-player multi-armed bandits. This setup allows us to model how players adapt and make decisions based on their historical observations.

4.1 Multi-Player Multi-Armed Bandits Framework

Inspired by previous work [Xu et al., 2023, Boursier and Perchet, 2022, Liu et al., 2020], we adopt a decentralized multi-player multi-armed bandit (MPMAB) framework where each 'arm' represents a resource. In this setting, players lack prior information regarding both the rewards offered by arms and their personal weights toward these arms[1]. Consequently, players must learn this information and maximize their payoffs through ongoing interactions with the platform.

Suppose there are T rounds. At each round $t \in [T]$, each arm (resource) k offers a stochastic reward $X_k(t)$, independently and identically distributed according to a distribution $\xi_k \in \Xi$, with expected value μ_k, supported on $(0, 1]$. We denote the set of all possible such distributions as Ξ. The vector of rewards from all arms at round t is represented as $\boldsymbol{X}(t) = \{X_k(t)\}_{k=1}^K$. During each round $t \in [T]$, N players select arms concurrently, based solely on their past observations. We use $\pi_j(t)$ to indicate the arm chosen by player j and $\boldsymbol{\pi}(t) = \{\pi_j(t)\}_{j=1}^N$ to denote the strategy profile of all players at that round.

When multiple players select the same arm, the reward is allocated according to the proportional payoff rule specified in Equation (1). Specifically, let $R_j(t)$ represent the reward received by player j at round t, given by:

$$R_j(t) = X_k(t) \cdot \frac{w_{j,k}}{\sum_{j'=1}^N \mathbb{I}[\pi_{j'}(t) = k] w_{j',k}}, \text{ where } \quad k = \pi_j(t). \tag{4}$$

Our framework aligns with the statistical sensing setting commonly found in MPMAB literature [Boursier and Perchet, 2022]. Specifically, each player can observe both their individual reward $R_j(t)$ and the total reward of the arm they have chosen, denoted as $X_{\pi_j(t)}(t)$.

In addition, consistent with previous works [Boursier and Perchet, 2020, Xu et al., 2023], our framework operates under the assumption of selfish players. In this setting, each player is focused on maximizing their own reward rather than collaborating to achieve a globally optimal outcome.

4.2 Evaluation metrics

We consider the following evaluation metrics.

Players' Regret. Since we consider the selfish player setting, we evaluate the regret of individual players by contrasting the expected reward garnered through their chosen strategy, $\boldsymbol{\pi}(t)$, with the best possible choice for each round t. The formal expression for player j's regret is as follows:

$$\mathrm{Reg}_j(T) = \sum_{t=1}^T \left(\max_{k \in [K]} \mathcal{U}_j((k, \boldsymbol{\pi}_{-j}(t)) - \mathcal{U}_j(\boldsymbol{\pi}(t)) \right). \tag{5}$$

Notably, $\mathrm{Reg}_j(T)$ is a random variable and the randomness come from the actions $\{\boldsymbol{\pi}(t)\}$, which relies on historical random observations $\{X_k(t)\}$.

Number of Non-inefficient PNE Rounds. As demonstrated in Section 3.3.3, the game can have multiple Pure Nash Equilibria (PNE), some of which may be suboptimal in terms of total welfare. Therefore, an important metric we consider is not merely the convergence to any PNE, but specifically to the most efficient PNE. We evaluate this using the number of rounds until round T in which players do not converge to the most efficient PNE. Mathematically, this is expressed as:

$$\mathrm{NonPNE}(T) = \sum_{t=1}^T \mathbb{I}[\boldsymbol{\pi}(t) \text{ is not a most efficient PNE of } \mathcal{G}]. \tag{6}$$

The metric $\mathrm{NonPNE}(T)$ provides a measure of how frequently the players end up in a suboptimal equilibrium.

[1] In the MPMAB framework, the terms 'payoff' and 'reward' are synonymous, as are 'resource' and 'arm'.

4.3 Algorithm

Algorithm 1 PPA-Game Online Learner

Input: Player rank j, Total number of rounds T, Phase length parameters c_1, c_2, c_3, η, Perturbation range Γ, Exploration probability ϵ
Exploration phase:
Set $S_{j,k} = 0$ for all $k \in [K]$
for $t \leftarrow 1$ to $c_1 K$ **do**
    Pull arm $\pi_j(t) \leftarrow (t + j) \mod K + 1$ and set $k \leftarrow \pi_j(t)$
    Get the arm's total reward $X_k(t)$
    Update the sum of rewards $S_{j,k} \leftarrow S_{j,k} + X_k(t)$
end for
Estimate the rewards of each arm $\hat{\mu}_{j,k} \leftarrow S_{j,k}/c_1$ for $k \in [K]$
Randomly sample perturbations $\gamma_{j,k} \sim \text{Uniform}(0, \Gamma)$ for each $k \in [K]$
Learning and exploitation phase:
$s \leftarrow 0$
while $t \leq T$ **do**
    Learning PNE subphase:
    $s \leftarrow s + 1$.
    Initialize the counter $Q_{j,k}^{(s)} \leftarrow 0$.
    Initialize the status $(m_j, a_j, u_j) \leftarrow (D, 1, 0)$
    for the next $c_2 s^\eta$ rounds **do**
        $t \leftarrow t + 1$
        Pull arm $\pi_j(t)$ according to Rule 4.1 and set $k \leftarrow \pi_j(t)$
        Get the arm's total reward $X_k(t)$ and player j's own reward $R_j(t)$
        Set $u'_j \leftarrow R_j(t)\hat{\mu}_{j,k}/X_k(t) + \gamma_{j,k}$
        Update the status (m_j, a_j, u_j) according to Rule 4.2
        if $m_j = C$ **then**
            $Q_{j,k}^{(s)} \leftarrow Q_{j,k}^{(s)} + 1$
        end if
    end for
    Exploitation subphase:
    Find $k \leftarrow \arg\max_{k'} \sum_{s'=1}^{s} Q_{j,k'}^{s'}$
    for the next $c_3 2^s$ rounds **do**
        $t \leftarrow t + 1$
        Pull arm $\pi_j(t) \leftarrow k'$
    end for
end while

Inspired by [Bistritz and Leshem, 2018, 2021, Pradelski and Young, 2012], we introduce a novel online learning algorithm designed to enable players to both learn the game's parameters and optimize their individual rewards concurrently. The algorithm's framework is shown in Figure 1 and pseudo-code is outlined in Algorithm 1.

Generally speaking, the algorithm operates in two distinct phases: the Exploration Phase and the Learning and Exploitation Phase. During the Exploration Phase, each player pulls each arm multiple times. This serves to estimate the arm rewards and facilitates the computation of the most efficient Pure Nash Equilibrium (PNE) for the subsequent phase. Furthermore, in the Learning and Exploitation Phase, players alternate between two subphases. The first subphase focuses on identifying the most efficient PNE, following the approach presented in [Pradelski and Young, 2012]. The second subphase centers on exploiting the gathered information to commit to the identified most efficient PNE.

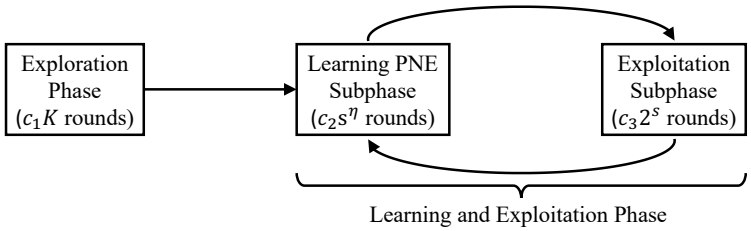

Figure 1: Graphical overview of our algorithm: Utilizes hyper-parameters c_1, c_2, c_3, and η, with counter s progressing through natural numbers. Initially, the algorithm enters an Exploration Phase for c_1K rounds (Section 4.3.1), followed by a Learning and Exploitation phase (Section 4.3.2), where it alternates between Learning PNE and Exploitation Subphases based on the counter s. Specifically, in round s, the Learning PNE Subphase spans $c_2 s^\eta$ rounds, and the Exploitation Subphase spans $c_3 2^s$ rounds.

4.3.1 Exploration Phase

During this phase, each player pulls each arm c_1 times to estimate the expected reward $\hat{\mu}_{j,k}$ for each arm k. This estimate is calculated as the average of the observed rewards when pulling the arm. To further reduce regret during this phase, a collision-minimizing strategy is employed: at each round t, player j selects the arm using the formula $(j+t) \mod K+1$. This strategy aims to minimize the number of collisions among players.

4.3.2 Learning and Exploitation Phase

We next move on to the Learning and Exploitation Phase, which cyclically alternates between the Learning PNE and Exploitation Subphases.

Although our ultimate goal is to identify the most efficient Nash equilibrium in the game $\mathcal{G}$ as outlined in Definition 3.2, precise parameter estimation for $\mathcal{G}$ is unfeasible. Players can only approximate the game based on their arm reward estimates $\hat{\mu}_{j,k}$. Additionally, due to the potential non-uniqueness of the most efficient Nash equilibria in $\mathcal{G}$, we introduce a randomized perturbation term $\gamma_{j,k}$ to each player's utility. Specifically, we aim to find the most efficient PNE in a perturbed game $\mathcal{G}'$, which is given as:

$$\begin{aligned} &\mathcal{G}' = \left(\mathcal{S}, \{\mathcal{U}'_j(\boldsymbol{\pi})\}_{j=1}^N\right), \quad \text{where} \\ &\mathcal{S} = [K] \text{ and } \mathcal{U}'_j(\boldsymbol{\pi}) = \frac{\hat{\mu}_{j,k} w_{j,k}}{\sum_{j'=1}^N \mathbb{I}[\pi_{j'} = k] w_{j',k}} + \gamma_{j,k}, k = \pi_j. \end{aligned} \tag{7}$$

Here, $\gamma_{j,k}$ is randomly sampled from a uniform distribution over the interval $[0, \Gamma]$. As proven in Theorem 4.1, under certain conditions, $\mathcal{G}'$ has a unique most efficient PNE with probability 1, which also serves as a most efficient PNE in the original game $\mathcal{G}$. Now we introduce the details of Learning PNE and Exploitation Subphases.

Learning PNE Subphase. In this subphase, we adopt the framework proposed in [Pradelski and Young, 2012] to identify the most efficient Nash equilibrium. Specifically, each player maintains a status tuple (m_j, a_j, u_j) at each round t during this phase. Here, m_j can be one of four moods: discontent (D), content (C), watchful (C^-), and hopeful (C^+). Additionally, a_j denotes the arm currently being played by player j, and u_j represents the utility player j gains when playing arm a_j.

Intuitively, the concept of mood shows a player's attitude towards their current strategy. A discontent player is uncertain about which arm is the most profitable, necessitating further exploration. Conversely, a content player believes they are already following the most efficient PNE. A watchful player is skeptical about their current strategy's efficacy and may transition to a discontent mood. Finally, a hopeful player believes that their current strategy will yield higher-than-expected returns, and thus continues to adhere to it.

At each round t, players pull arms in accordance with Rule 4.1 based on their mood from the previous round.

Rule 4.1 (Rule to pull arm)**.** The chosen arm $\pi_j(t)$ is chosen based on the mood at the last round.

- A discontent (D) player chooses arm $\pi_j(t)$ uniformly at random from $[K]$.
- A content (C) player chooses arm $\pi_j(t) = a_j$ with probability $1 - \epsilon$ and chooses all other arms with probability $\epsilon/(K-1)$.
- A watchful (C^-) or hopeful (C^+) player chooses arm $\pi_j(t) = a_j$.

Remark 4.1. Rule 4.1 captures the essence of the different moods. Specifically, discontent players explore the arm space uniformly, while content players predominantly stick to their current strategy but also explore other strategies with low probability. Watchful and hopeful players require additional rounds to ascertain the efficacy of their current strategy, so they maintain their current choice of arm.

After pulling the selected arm $\pi_j(t)$, player j observes his own reward $R_j(t)$ and the total reward $X_k(t)$ for the arm, where $k = \pi_j(t)$. The utility is then calculated as

$$u'_j = R_j(t)\hat{\mu}_{j,k}/X_k(t) + \gamma_{j,k}. \tag{8}$$

The player's status is updated based on this computed utility u'_j, in accordance with Rule 4.2.

Rule 4.2 (Rule to update status)**.** Define the functions $F(u)$ and $G(u)$ as follows:

$$F(u) = -\frac{u}{(4+3\Gamma)N} + \frac{1}{3N} \quad \text{and} \quad G(u) = -\frac{u}{4+3\Gamma} + \frac{1}{3}. \tag{9}$$

The status update is outlined below:

- A discontent (D) player transits to status $(m_j, a_j, u_j) \leftarrow (C,\ \pi_j(t), u'_j)$ with probability $\epsilon^{F(u'_j)}$ and remains unchanged otherwise.
- A content (C) player's rule is based on his own choice $\pi_j(t)$. If $\pi_j(t) = a_j$, we have

$$(m_j, a_j, u_j) \leftarrow \begin{cases} (C^+, a_j, u_j) & \text{if } u'_j > u_j, \\ (C, a_j, u_j) & \text{if } u'_j = u_j, \\ (C^-, a_j, u_j) & \text{if } u'_j < u_j. \end{cases} \tag{10}$$

  And under the case when $\pi_j(t) \neq a_j$, we have that if $u'_j > u_j$ then $(m_j, a_j, u_j) \leftarrow (C, \pi_j(t), u'_j)$ with probability $\epsilon^{G(u'_j - u_j)}$ and remains unchanged with probability $1 - \epsilon^{G(u'_j - u_j)}$. If $u'_j \leq u_j$, (m_j, a_j, u_j) remains unchanged.

- A watchful (C^-) player follows

$$(m_j, a_j, u_j) \leftarrow \begin{cases} (C^+, a_j, u_j) & \text{if } u'_j > u_j, \\ (C, a_j, u_j) & \text{if } u'_j = u_j, \\ (D, a_j, u_j) & \text{if } u'_j < u_j. \end{cases} \tag{11}$$

- A hopeful (C^+) player follows

$$(m_j, a_j, u_j) \leftarrow \begin{cases} (C, a_j, u'_j) & \text{if } u'_j > u_j, \\ (C, a_j, u_j) & \text{if } u'_j = u_j, \\ (C^-, a_j, u_j) & \text{if } u'_j < u_j. \end{cases} \tag{12}$$

Remark 4.2. Rule 4.2 is designed to reflect the intuitions behind the various moods. For discontent players, the status transitions to content if the randomly chosen strategy yields high utility. Content players, on the other hand, adjust their status based on the observed utility at the current round if they have not explored other strategies; otherwise, they transition to a different strategy based on utility. Watchful and hopeful players adjust their status based on the observed utility as well.

In line with existing work [Pradelski and Young, 2012, Young, 2009, Marden et al., 2014], Rule 4.1 and 4.2 constitute a trial-and-error learning procedure, akin to common learning routines in human behavior [Young, 2009].

Additionally, during the Learning PNE Subphase, each player maintains a counter $Q_{j,k}$ for each arm $k \in [K]$. This counter records the number of times that arm k has been chosen while the player is in a Content mood.

Exploitation Subphase. During the Exploitation Subphases, each player chooses the arm with the maximal number of times that arm is chosen with a content mood, *i.e.*, $k = \arg\max_{k'} Q_{j,k'}$. Then the player will pull arm k for the remaining rounds in the subphase.

4.4 Theoretical Analysis

In this subsection, we analyze the performance of our online algorithm. We first establish a metric to quantify the challenge of identifying the most efficient PNE in the presence of imprecise resource payoffs. Subsequently, this metric serves as a basis for deriving rigorous bounds on both regret and the number of rounds in which players do not follow the most efficient PNE.

4.4.1 Quantifying the Challenge of Identifying the Most Efficient PNE

We introduce the sets A^{NE} and A^{NoNE} to denote Nash equilibria and non-equilibria strategy profiles, respectively. Let

$$\begin{aligned} \delta^{\mathrm{NE}} &= \min\left\{\mathcal{U}_j(\boldsymbol{\pi}) - \mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) : \boldsymbol{\pi} \in A^{\mathrm{NE}}, j \in [N], \pi' \in [K], \pi' \neq \pi_j\right\}. \\ \delta^{\mathrm{NoNE}} &= \sup\left\{\delta' : \forall \boldsymbol{\pi} \in A^{\mathrm{NoNE}}, \boldsymbol{\pi} \text{ is not a } \delta'\text{-Nash equilibrium}\right\}. \\ \delta &= \min\left\{\delta^{\mathrm{NE}}, \delta^{\mathrm{NoNE}}\right\}. \end{aligned} \tag{13}$$

Here, δ^{NE} quantifies the probability of erroneously disregarding a PNE π due to inaccuracies in reward estimation. Conversely, δ^{NoNE} gauges the risk of misidentifying a non-equilibrium strategy profile π as a PNE. To capture both of these risks, we set δ as the minimum of δ^{NE} and δ^{NoNE}. We further show from the following example that both δ^{NE} and δ^{NoNE} may determine δ.

Example 4.1. Set $N = 3$ and $K = 2$. Here we suppose that each player has the same weights on the resources (as per Condition (2) in Theorem 3.2). When arms' rewards are $[1, 0.7]$ and players' weights are $[1, 0.8, 0.4]$, we have $0.052 = \delta^{\mathrm{NE}} > \delta^{\mathrm{NoNE}} = 0.026$. By contrast, when arms' rewards are $[1, 0.6]$ and players' weights are $[1, 2/3, 4/9]$, we have $0.040 = \delta^{\mathrm{NE}} < \delta^{\mathrm{NoNE}} = 0.068$. Details of the two cases can refer to Tables 1 and 2 in Appendix B.

Let Δ be the largest discrepancy between the estimated and true rewards $\Delta = \max_{j,k} |\hat{\mu}_{j,k} - \mu_k|$. We can have the following theorem.

Theorem 4.1. *Suppose $\Delta < \delta/(4K)$, $\Gamma \leq \delta/(4N)$, and the PNE exists for the game $\mathcal{G}$ in Definition 3.2. Then the PNE exists for the game $\mathcal{G}'$ (Equation (7)). In addition, the most efficient PNE in the game $\mathcal{G}'$ is unique with probability 1 and the equilibrium is a most efficient PNE in the original game $\mathcal{G}$.*

Remark 4.3. The proof is provided in Appendix C.3. Theorem 4.1 demonstrates that under bounded estimation error Δ and perturbation Γ, the most efficient PNE in $\mathcal{G}$ can be uniquely identified.

4.4.2 Performance Analysis

We now analyze the two metrics introduced in Section 4.2. Following [Bistritz and Leshem, 2018, 2021], the Learning PNE Subphase defines a Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$ on the states of all players $Z = \{(m_j, a_j, u_j)\}_{j\in[N]}$ based on the parameter ϵ and the game $\mathcal{G}'$. Based on this observation, we could obtain the following results. Note that c_1, c_2, c_3, η are hyper-parameters of our algorithm as shown in Figure 1.

Theorem 4.2. *Suppose $\Delta < \delta/(4K)$, $\Gamma \leq \delta/(4N)$, $w_{j,k} > 0$ for all $j \in [N]$ and $k \in [K]$, and the PNE exists for the game $\mathcal{G}$ in Definition 3.2. Set $c_1 \geq 8K^2 \log T/\delta^2$. There exists a small enough $\epsilon > 0$ such that when T is large enough, we have that $\forall j \in [N]$,*

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}_j(T)] &\leq O\left(\mu_{\max}\left(c_1 + c_2 \log^{1+\eta} T + c_3 \log T\right)\right), \\ \mathbb{E}[\mathrm{NonPNE}(T)] &\leq O\left(c_1 + c_2 \log^{1+\eta} T + c_3 \log T\right), \end{aligned} \tag{14}$$

where $\mu_{\max} = \max \mu_k$.

Remark 4.4. The proof of Theorem 4.2 is detailed in Appendix C.4. The theorem establishes that our algorithm converges to the most efficient PNE in the long term. It further characterizes the regret for each player as $O(\log^{1+\eta} T)$ for any $\eta > 0$, under sufficient conditions. This aligns with existing results in [Bistritz and Leshem, 2021]. Additionally, as $\eta \to 0$, our bounds reduce to $O(\log T)$, consistent with standard regret bounds in the multi-armed bandit literature [Boursier and Perchet, 2022, Slivkins et al., 2019].

5 Synthetic Experiments

In this section, we validate the effectiveness of our online method through synthetic experiments.

5.1 Data-generating Process

In our experiments, we assess two distinct scenarios: $(N, K) = (3, 5)$ and $(N, K) = (5, 3)$, effectively examining both $N < K$ and $N > K$ cases. We set the time horizon, T, to 3,000,000 for each scenario. For modeling the payoff distribution of each resource, we use the beta distribution, constrained within the interval $[0, 1]$. Specifically, for each resource k, we independently draw two parameters, α_k and β_k, from a uniform distribution over $[0, 10]$. The rewards for each resource (or arm) in different rounds are then sampled from $\xi_k = \text{Beta}(\alpha_k, \beta_k)$. The players' weights $w_{j,k}$ are drawn from $\text{Uniform}(0, 1)$.

We further explore two specific scenarios in the context of players' weights: (1) *Homogeneous setting*, where all players have identical weights across various resources, aligning with the second condition in Theorem 3.2; and (2) *Standard setting*, where players' weights across resources can differ without constraints.

5.2 Baselines and Hyper-parameters

We benchmark our approach against the following baselines.

- **TotalReward**: Proposed by Bande and Veeravalli [2019], this method emphasizes a shareable reward case, aiming to maximize the collective reward for all players. In scenarios where $N < K$, the optimal strategy for each round is to have all players select the top N arms. When $N \geq K$, every arm should be selected in every round. This strategy operates in an explore-then-commit manner. Specifically, during the initial $\alpha \log T$ rounds, the algorithm explores arms randomly. Thereafter, it commits to the optimal strategy in each round. The hyper-parameter α is chosen from the set $\{100, 200, 500, 1000, 2000\}$.

- **SelfishRobustMMAB**: Proposed by Boursier and Perchet [2020], this algorithm aims to find the most efficient PNE. However, it assumes that when multiple players select the same resource, none receive a payoff. Under this context, the method provides algorithms resistant to the deviations of individual selfish players. The algorithm utilizes KL-UCB and a coefficient β to determine the exploration of sub-optimal arms. The hyper-parameter β is examined over the range $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

- **SMAA**: According to Xu et al. [2023], this approach also aims to identify the PNE. It assumes an averaging allocation model, meaning players equally divide the payoffs of resources when selecting the same one. Similar to [Boursier and Perchet, 2020], a hyper-parameter β is utilized to modulate the exploration of sub-optimal arms. We explore values for β in the set $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

As for our method, we slightly abuse the notation for clarity. During the exploration phase, each arm is explored for a duration of $c_1 K^2 \log T / \delta^2$. The value of Γ is set to $\delta/(4N)$ and c_3 is set to 200. Our search encompasses various hyper-parameters, including c_1 from $\{0.001, 0.01\}$, c_2 from $\{1000, 5000\}$, η from $\{1, 2\}$, and ϵ from $\{0.001, 0.003, 0.005\}$.

5.3 Analysis

We conducted experiments across 50 unique simulations by repeatedly sampling the resources' payoff distributions and players' weights. Our method, alongside the baseline approaches, was evaluated by computing

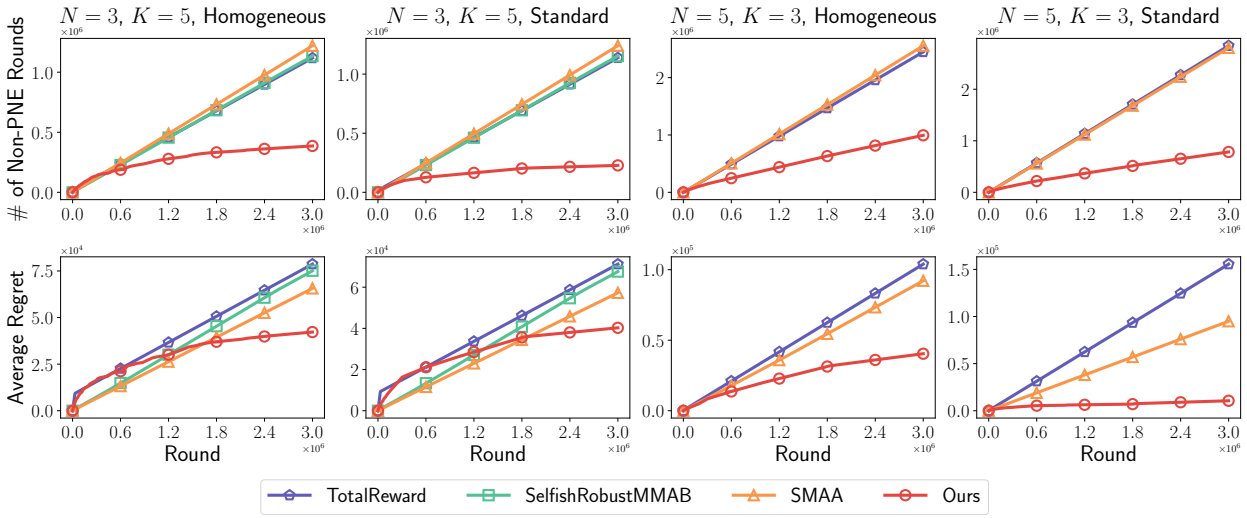

Figure 2: Curves showing the average regret and the number of rounds where players do not follow the most efficient PNE. Note that SelfishRobustMMAB could not be applied to scenarios when $N > K$.

the average regret across all players and quantifying the number of rounds where players diverged from the most efficient PNE. The experimental results are depicted in Figure 2.

From the figure, we can observe that our methodology consistently surpasses all baselines across different data-generating processes. Given that none of the baselines is designed specifically for our PPA-Game, they often fail to identify the PNE. Consequently, their performance curves are linear across all setups, indicating that both regrets and the count of Non-PNE rounds increase linearly with T. By contrast, our method exhibits a diminishing slope as T grows, reinforcing the idea that our approach eventually converges to the most efficient PNE. These experimental findings also demonstrate that the PNE solution can vary significantly based on game formulations, emphasizing the pivotal role of precise game formulation in practical scenarios.

6 Conclusions

In summary, we introduce the Proportional Payoff Allocation Game (PPA-Game) as a novel framework for analyzing online content creator competition, confirming the frequent emergence of pure Nash equilibrium (PNE) through extensive simulations. Further, our study extends to the Multi-player Multi-Armed Bandit (MPMAB) framework, exploring creators' adaptive strategies in dynamic demand environments, and validates an effective new online algorithm for content optimization, backed by theoretical and experimental analysis.

Acknowledgements

Peng Cui's research was supported in part by National Natural Science Foundation of China (No. 62141607). 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.

References

Nejat Anbarci, Kutay Cingiz, and Mehmet S Ismail. Proportional resource allocation in dynamic n-player blotto games. *Mathematical Social Sciences*, 125:94–100, 2023.

Ayala Arad and Ariel Rubinstein. Multi-dimensional iterative reasoning in action: The case of the colonel blotto game. *Journal of Economic Behavior & Organization*, 84(2):571–585, 2012.

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

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

Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

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

Omer Ben-Porat, Gregory Goren, Itay Rosenberg, and Moshe Tennenholtz. From recommendation systems to facility location games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1772–1779, 2019a.

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

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

Mohammed Anis Benblidia, Bouziane Brik, Moez Esseghir, and Leila Merghem-Boulahia. A renewable energy-aware power allocation for cloud data centers: A game theory approach. *Computer Communications*, 179:102–111, 2021.

Lilian Besson and Emilie Kaufmann. Multi-player bandits revisited. In *Algorithmic Learning Theory*, pages 56–92. PMLR, 2018.

Kshipra Bhawalkar, Martin Gairing, and Tim Roughgarden. Weighted congestion games: the price of anarchy, universal worst-case examples, and tightness. *ACM Transactions on Economics and Computation (TEAC)*, 2(4):1–23, 2014.

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

Ilai Bistritz and Amir Leshem. Game of thrones: Fully distributed learning for multiplayer bandits. *Mathematics of Operations Research*, 46(1):159–178, 2021.

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

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

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

Craig Boutilier, Martin Mladenov, and Guy Tennenholtz. Modeling recommender ecosystems: Research challenges at the intersection of mechanism design, reinforcement learning and generative models. *arXiv preprint arXiv:2309.06375*, 2023.

Tomer Boyarski, Amir Leshem, and Vikram Krishnamurthy. Distributed learning in congested environments with partial information. *arXiv preprint arXiv:2103.15901*, 2021.

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

Ioannis Caragiannis and Alexandros A Voudouris. Welfare guarantees for proportional allocations. *Theory of Computing Systems*, 59:581–599, 2016.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. In *29th International Symposium on Theoretical Aspects of Computer Science*, page 124, 2012.

Lawrence Friedman. Game-theory models in the allocation of advertising expenditures. *Operations research*, 6(5):699–709, 1958.

Martin Hoefer and Wanchote Jiamjitrak. On proportional allocation in hedonic games. In *International Symposium on Algorithmic Game Theory*, pages 307–319. Springer, 2017.

Jiri Hron, Karl Krauth, Michael I Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. *arXiv preprint arXiv:2206.13102*, 2022.

Wei Huang, Richard Combes, and Cindy Trinh. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Conference on Learning Theory*, pages 1990–2012. PMLR, 2022.

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34: 3323–3335, 2021.

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. *arXiv preprint arXiv:2206.13489*, 2022.

Ramesh Johari. The price of anarchy and the design of scalable resource allocation mechanisms. *Algorithmic Game Theory*, pages 543–568, 2007.

Frank Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8 (1):33–37, 1997.

Tilman Klumpp, Kai A Konrad, and Adam Solomon. The dynamics of majoritarian blotto games. *Games and Economic Behavior*, 117:402–419, 2019.

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

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

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22:1–34, 2021.

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

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

Jason R Marden, H Peyton Young, and Lucy Y Pao. Achieving pareto optimality through distributed learning. *SIAM Journal on Control and Optimization*, 52(5):2753–2770, 2014.

Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard Zemel, and Craig Boutilier. Optimizing long-term social welfare in recommender systems: A constrained matching approach. In *International Conference on Machine Learning*, pages 6987–6998. PMLR, 2020.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

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

Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference 2020*, pages 1194–1204, 2020.

Bary SR Pradelski and H Peyton Young. Learning efficient nash equilibria in distributed systems. *Games and Economic behavior*, 75(2):882–897, 2012.

Siddharth Prasad, Martin Mladenov, and Craig Boutilier. Content prompting: Modeling content provider dynamics to improve user welfare in recommender ecosystems. *arXiv preprint arXiv:2309.00940*, 2023.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 155–163. PMLR, 2016.

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

Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Decentralized multi-player multi-armed bandits with no collision information. In *International Conference on Artificial Intelligence and Statistics*, pages 1519–1528. PMLR, 2020.

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

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

Statista. Most popular categories on tiktok worldwide as of 2023, by hashtag views, 2023. URL https://www.statista.com/statistics/1130988/most-popular-categories-tiktok-worldwide-hashtag-views/. Accessed: October 13, 2023.

Vasilis Syrgkanis and Eva Tardos. Composable and efficient mechanisms. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 211–220, 2013.

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

Xuchuang Wang, Hong Xie, and John CS Lui. Multiple-play stochastic bandits with shareable finite-capacity arms. In *International Conference on Machine Learning*, pages 23181–23212. PMLR, 2022b.

Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, and Peng Cui. Competing for shareable arms in multi-player multi-armed bandits. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38674–38706. PMLR, 2023.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? *arXiv preprint arXiv:2302.01971*, 2023.

H Peyton Young. The evolution of conventions. *Econometrica: Journal of the Econometric Society*, pages 57–84, 1993.

H Peyton Young. Learning by trial and error. *Games and economic behavior*, 65(2):626–643, 2009.

Marie-Joséph Youssef, Venugopal V Veeravalli, Joumana Farah, Charbel Abdel Nour, and Catherine Douillard. Resource allocation in noma-based self-organizing networks using stochastic multi-armed bandits. *IEEE Transactions on Communications*, 69(9):6003–6017, 2021.

Ruohan Zhan, Konstantina Christakopoulou, Ya Le, Jayden Ooi, Martin Mladenov, Alex Beutel, Craig Boutilier, Ed Chi, and Minmin Chen. Towards content provider aware recommender systems: A simulation study on the interplay between user and provider utilities. In *Proceedings of the Web Conference 2021*, pages 3872–3883, 2021.

A Further Analysis on Pure Nash Equilibrium (PNE)

A.1 The Non-existence of PNE in Certain Scenarios

To show the fact that a Pure Nash Equilibrium (PNE) might not always exist, we present a specific counter-example below.

Example A.1. Let's consider a configuration with $N = 7$ players and $K = 4$ resources. Assign the resources' payoffs as $\boldsymbol{\mu} = (1, 0.4904, 0.2447, 0.1065)$. The matrix representing the weights of the players for each resource is given by Equation (15). A comprehensive enumeration of possible strategy profiles confirms the non-existence of any PNE in this particular configuration.

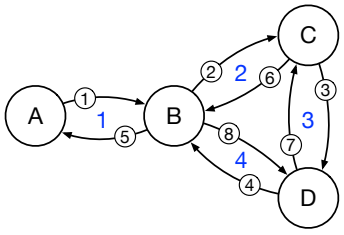

Figure 3: A showcase in which PNE does not exist.

$$\boldsymbol{W} = \begin{pmatrix} 0.1657 & 0 & 0.5755 & 0 \\ 0 & 0.0314 & 0 & 0.5623 \\ 0 & 0 & 0.4217 & 0.9517 \\ 0 & 1 & 0.4057 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{15}$$

We find an improvement cycle for players 1, 2, 3, and 4 in this example, as depicted in Figure 3. In this figure, the large circles labeled A, B, C, and D represent resources. The blue numbers between these resource circles signify the players and the sequence of their actions, while the numbers inside the larger circles indicate the respective rounds. At round 1, player 1 perceives that resource B offers a more favorable payoff and hence shifts to resource B. This transition decreases the payoff for player 2, prompting him to switch to resource C by round 2. The cycle continues, and by round 4, player 4 transitions from resource D to B, which subsequently reduces the payoff for player 1. This leads player 1 to revert to resource A by round 5, and in turn, prompts player 2 to shift back from resource C to B by round 6, thus perpetuating the cycle.

A.2 High Probability of PNE Existence

We experiment with cases where the values of N and K are large, using Monte Carlo sampling. We randomly select values for N and K, uniformly from $[10, 100]$. Moreover, for each player-resource pair (j, k), with $j \in [N]$ and $k \in [K]$, weights $w_{j,k}$ and payoffs μ_k are sampled from four different types of distributions.

- **Uniform Distribution**: We sample payoffs and weights using a uniform distribution $U(0, 1)$.
- **Truncated Normal Distribution**: We sample payoffs and weights using a normal distribution $N(0.5, 0.1^2)$ and truncate them into $[0, 1]$. This distribution provides a scenario where values are concentrated around the mean of 0.5.
- **Truncated T-Distribution**: We utilize a t-distribution with 5 degrees of freedom, adjusted by adding 0.5 to center the mean at $[0, 1]$. Similarly, we truncate the values into $[0, 1]$. This distribution also provides a scenario where values are concentrated around the mean of 0.5.
- **Beta Distribution**: We opt for a beta distribution with parameters $\alpha = \beta = 0.5$. This distribution is skewed towards the extremes of 0 and 1, representing scenarios where values are highly polarized.

For every randomly generated game configuration, considering the computational complexity of evaluating all strategy profiles, we trace an improvement path starting from the strategy profile $\boldsymbol{\pi}(0) = (0, 0, \ldots, 0)$. If this improvement path concludes within finite steps, it implies the existence of a PNE. Out of the $5,000,000$ randomly generated game configurations in each distribution (*i.e.*, totaling $20,000,000$ randomly generated

configurations), only one exhibited an infinite improvement path. This suggests the existence of a PNE in the remaining $19,999,999$ configurations. A comprehensive evaluation of strategy profiles confirmed the non-existence of a PNE in the only exceptional case, concisely illustrated in Appendix A.1.

B Details About Example 4.1

The details of toy cases in Example 4.1 can be found in Tables 1 and 2.

Table 1: A showcase that $\delta^{\mathrm{NE}} > \delta^{\mathrm{NoNE}}$. Set $N = 3$ and $K = 2$. Assume the resources' payoffs to be $[1, 0.7]$. Suppose each player has the same weights on the resources (as per Condition (2) in Theorem 3.2), with weights $[1, 0.8, 0.4]$, respectively. We list all strategy profiles and find that $\delta^{\mathrm{NE}} = \min\{0.052, 0.133\} = 0.052 > 0.022 = \min\{0.518, 0.022, 0.052, 0.133, 0.356, 0.873\} = \delta^{\mathrm{NoNE}}$.

Strategy profile	PNE?	Least reward change by deviation if PNE	δ'-Nash equilibrium if not PNE
$(1,1,1)$	✗		0.518. Deviated strategy $(1,1,2)$
$(1,1,2)$	✗		0.022. Deviated strategy $(1,2,2)$
$(1,2,1)$	✓	0.052. Deviated strategy $(1,2,2)$	
$(1,2,2)$	✗		0.052. Deviated strategy $(1,2,1)$
$(2,1,1)$	✓	0.133. Deviated strategy $(2,1,2)$	
$(2,1,2)$	✗		0.133. Deviated strategy $(2,1,1)$
$(2,2,1)$	✗		0.356. Deviated strategy $(2,1,1)$
$(2,2,2)$	✗		0.873. Deviated strategy $(2,2,1)$

Table 2: A showcase that $\delta^{\mathrm{NE}} < \delta^{\mathrm{NoNE}}$. Set $N = 3$ and $K = 2$. Assume the resources' payoffs to be $[1, 0.6]$. Suppose each player has the same weights on the resources (as per Condition (2) in Theorem 3.2), with weights $[1, 2/3, 4/9]$, respectively. We list all strategy profiles and find that $\delta^{\mathrm{NE}} = \min\{0.040, 0.068, 0.126\} = 0.040 < 0.068 = \min\{0.389, 0.068, 0.215, 0.360, 0.874\} = \delta^{\mathrm{NoNE}}$.

Strategy profile	PNE?	Least reward change by deviation if PNE	δ'-Nash equilibrium if not PNE
$(1,1,1)$	✗		0.389. Deviated strategy $(1,1,2)$
$(1,1,2)$	✓	0.040. Deviated strategy $(1,2,2)$	
$(1,2,1)$	✓	0.068. Deviated strategy $(1,2,2)$	
$(1,2,2)$	✗		0.068. Deviated strategy $(1,2,1)$
$(2,1,1)$	✓	0.126. Deviated strategy $(1,1,1)$	
$(2,1,2)$	✗		0.215. Deviated strategy $(2,1,1)$
$(2,2,1)$	✗		0.360. Deviated strategy $(2,1,1)$
$(2,2,2)$	✗		0.874. Deviated strategy $(2,2,1)$

C Omitted Proofs

C.1 Proof of Theorem 3.1

We first introduce some notations. For a strategy profile $\boldsymbol{\pi}$, define $W_k(\boldsymbol{\pi})$ as the sum of weights of players that choose resource k, *i.e.*,

$$W_k(\boldsymbol{\pi}) = \sum_{j=1}^{N} \mathbb{I}[\pi_j = k] w_{j,k}. \tag{16}$$

Furthermore, Let $\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1))$ be an improvement cycle. For any s such that $\boldsymbol{\pi}(s+1)$ is an improvement step of $\boldsymbol{\pi}(s)$, define $j(s)$ be the deviated player in this step and the new strategy

is $k(s)$, *i.e.*, $\boldsymbol{\pi}(s+1) = \big(k(s), \boldsymbol{\pi}(s)_{-j(s)}\big)$. Besides, we let $Q_k(\ell)$ be the minimal sum of players' weights on the resource k in the improvement cycle ℓ, *i.e.*,

$$Q_k(\ell) = \min_{\boldsymbol{\pi}(s)\in\ell} \sum_{j\in B_k(\boldsymbol{\pi}(s))} w_{j,k} = \min_{\boldsymbol{\pi}(s)\in\ell} W_k(\boldsymbol{\pi}(s)). \tag{17}$$

We first introduce the following lemmas.

Lemma C.1. *Let $\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1))$ be an improvement cycle. For any $1 \le s \le t$, let $j = j(s)$ and $k = k(s)$. Then we have*

$$\mathcal{U}_j(\boldsymbol{\pi}(s+1)) \le \frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}}. \tag{18}$$

Proof. Since $\boldsymbol{\pi}(s+1)$ is an improvement step of $\boldsymbol{\pi}(s)$, player j earns a larger payoff by deviation in this step, leading to the fact that $w_{j,k} > 0$ and the right-hand side of Equation (18) is well-defined. Furthermore, by the definition of $Q_k(\ell)$, we have

$$W_k(\boldsymbol{\pi}(s+1)) = W_k(\boldsymbol{\pi}(s)) + w_{j,k} \ge Q_k(\ell) + w_{j,k}. \tag{19}$$

As a result,

$$\mathcal{U}_j(\boldsymbol{\pi}(s+1)) = \frac{\mu_k w_{j,k}}{W_k(\boldsymbol{\pi}(s+1))} \le \frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}}. \tag{20}$$

Now the claim follows. □

Lemma C.2. *Let $\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1))$ be an improvement cycle. Let k be any resource such that $\exists 1 \le s' \le t$, $k(s') = k$. Then there exist an index s such that $\pi_{j(s)}(s) = k$ and*

$$\mathcal{U}_j\left(\boldsymbol{\pi}(s)\right) = \frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}} \quad \textit{where} \quad j = j(s). \tag{21}$$

Proof. Since $\exists 1 \le s' \le t$ such that $k(s') = k$, there must exists rounds when a player deviate from the resource k. As a result, there must exist a constant $1 \le s \le t$ such that

$$\pi_{j(s)}(s) = k \quad \text{and} \quad Q_k(\ell) = W_k(\boldsymbol{\pi}(s+1)). \tag{22}$$

As a result, let $j = j(s)$ and we have

$$\mathcal{U}_j(\boldsymbol{\pi}(s)) = \frac{\mu_k w_{j,k}}{W_k(\boldsymbol{\pi}(s))} = \frac{\mu_k w_{j,k}}{W_k(\boldsymbol{\pi}(s+1)) + w_{j,k}} = \frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}}. \tag{23}$$

Now the claim follows. □

Lemma C.3. *Let $\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1))$ be an improvement cycle. Then for every improvement step $1 \le s \le t$ and resource k such that $k(s) = k$, there exist $1 \le s' \le t$ and $k' = k(s')$ such that*

$$\frac{\mu_k}{Q_k(\ell) + \epsilon_0} < \frac{\mu_{k'}}{Q_{k'}(\ell) + \epsilon_0}. \tag{24}$$

Proof. By Lemma C.2, we have that there exists an index s' such that $j = j(s')$, $\pi_j(s') = k$, and

$$\mathcal{U}_j(\boldsymbol{\pi}(s')) = \frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}}. \tag{25}$$

Set $k' = k(s')$. Since $\boldsymbol{\pi}(s'+1)$ is an improvement step of $\boldsymbol{\pi}(s')$ for player $j = j(s')$, we have

$$\frac{\mu_k w_{j,k}}{Q_k(\ell) + w_{j,k}} = \mathcal{U}_j(\boldsymbol{\pi}(s')) < \mathcal{U}_j(\boldsymbol{\pi}(s'+1)) \le \frac{\mu_{k'} w_{j,k'}}{Q_{k'}(\ell) + w_{j,k'}}, \tag{26}$$

where the last inequality is due to Lemma C.1.

First, we claim that $\mu_k \le \mu_{k'}$ by contradiction. Otherwise, we assume $\mu_k > \mu_{k'}$ and we obtain that $\mu_k/\mu_{k'} > N_0/\epsilon_0$ by the assumption. Based on Equation (26), we have:

$$\frac{\mu_{k'}}{\mu_k} > \frac{w_{j,k}}{Q_k(\ell) + w_{j,k}} \Big/ \frac{w_{j,k'}}{Q_{k'}(\ell) + w_{j,k'}} \,. \tag{27}$$

According to the assumption of ϵ_0 and N_0, and acknowledging that the function $\frac{x}{x+Q}$ is increasing, we have:

$$\frac{w_{j,k}}{Q_k(\ell) + w_{j,k}} \ge \frac{\epsilon_0}{N_0} \quad \text{and} \quad \frac{w_{j,k'}}{Q_{k'}(\ell) + w_{j,k'}} \le 1. \tag{28}$$

As a result,

$$\frac{\mu_{k'}}{\mu_k} > \frac{w_{j,k}}{Q_k(\ell) + w_{j,k}} \Big/ \frac{w_{j,k'}}{Q_{k'}(\ell) + w_{j,k'}} \ge \frac{\epsilon_0}{N_0} > \frac{\mu_{k'}}{\mu_k}. \tag{29}$$

This leads to a contradiction. As a result, we have $\mu_k \le \mu_{k'}$. Hence, by the assumption, we have

$$\frac{\mu_{k'}}{\mu_k} > \frac{N_0}{\epsilon_0} \ge \frac{W_{k'}(\boldsymbol{\pi}(s'+1))}{Q_k(\ell) + \epsilon_0} \ge \frac{Q_{k'}(\ell) + \epsilon_0}{Q_k(\ell) + \epsilon_0}. \tag{30}$$

Now the claim follows. □

With these lemmas, we can prove the original theorem.

Proof of Theorem 3.1. Suppose there exists an infinite improvement path. Then there must exist an improvement cycle

$$\ell = (\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1)) \,. \tag{31}$$

Let $s_1 = 1$ and $k_1 = k(s_1)$. According to Lemma C.3, there exist $1 \le s_2 \le t$ and $k_2 = k(s_2)$ such that $\mu_{k_1}/(Q_{k_1}(\ell) + \epsilon_0) < \mu_{k_2}/(Q_{k_2}(\ell) + \epsilon_0)$. Continue this process for $t+1$ times and we have $s_1, s_2, \ldots, s_{t+1}$ and $k_1, k_2, \ldots, k_{t+1}$ with $k_i = k(s_i)$ such that

$$\frac{\mu_{k_1}}{Q_{k_1}(\ell) + \epsilon_0} < \frac{\mu_{k_2}}{Q_{k_2}(\ell) + \epsilon_0} < \cdots < \frac{\mu_{k_{t+1}}}{Q_{k_{t+1}}(\ell) + \epsilon_0}. \tag{32}$$

Since k_i $(1 \le i \le t+1)$ can only take value in $\{1, 2, \ldots, t\}$, there must exists two indices $i_1 \ne i_2$ such that $k_{i_1} = k_{i_2}$. This leads to a contradiction since Equation (32) suggests that $\forall 1 \le i_1, i_2 \le t+1$, $k_{i_1} \ne k_{i_2}$. Now we can conclude that all improvement paths end in finite steps and the PNE exists. □

C.2 Proof of Theorem 3.2

We share similar proof techniques for the three conditions. Generally speaking, we want to prove that every improvement path is finite, which leads to the existence of PNE. In addition, for a strategy profile $\boldsymbol{\pi}$, define $W_k(\boldsymbol{\pi})$ as the sum of weights of players that choose resource k, *i.e.*,

$$W_k(\boldsymbol{\pi}) = \sum_{j=1}^{N} \mathbb{I}[\pi_j = k] w_{j,k}. \tag{33}$$

For any s such that $\boldsymbol{\pi}(s+1)$ is an improvement step of $\boldsymbol{\pi}(s)$, define $j(s)$ be the deviated player in this step and the new strategy is $k(s)$, *i.e.*, $\boldsymbol{\pi}(s+1) = \big(k(s), \boldsymbol{\pi}(s)_{-j(s)}\big)$.

Proof of Condition (1). Suppose there exists an infinite improvement path. Then there must exist an improvement cycle

$$(\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1)) \,. \tag{34}$$

Since for all $j \in [N]$ and $k, k' \in [K]$, we have $w_{j,k} = w_{j,k'}$, we slightly abuse the notation here and use w_j to denote $w_{j,k}$ for all $k \in [K]$. In addition, for any $1 \le s \le t$, define the potential function as

$$\Phi(\boldsymbol{\pi}(s)) = \mathrm{Sort}\left(\frac{\mu_k}{W_k(\boldsymbol{\pi}(s))}\right)_{k=1}^{K}. \tag{35}$$

Here $\Phi(\boldsymbol{\pi}(s))$ is a sorted vector and each element in the vector corresponds to a unit reward in a resource. Note that when $W_k(\boldsymbol{\pi}(s)) = \sum_{j=1}^{N} \mathbb{I}[\pi_j(s) = k]w_j = 0$ for some s and k, we set the value $\mu_k/W_k(\boldsymbol{\pi}(s))$ be infinity (∞). We suppose that $a < \infty$ for any real number a and $\infty = \infty$.

For each $1 \le s \le t$, since $\boldsymbol{\pi}(s+1) = \big(k(s), \boldsymbol{\pi}_{-j(s)}(s)\big)$ is an improvement step of $\boldsymbol{\pi}(s)$ and $\mu_{\pi_{j(s)}(s)} = \mu_{k(s)}$ by condition, setting $j = j(s)$ for brevity, we have

$$\frac{\mu_{\pi_j(s+1)}}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))} > \frac{\mu_{\pi_j(s)}}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))}. \tag{36}$$

In addition, note that player j deviate from resource $\pi_j(s)$ to $\pi_j(s+1)$ and we have that $W_{\pi_j(s)}(\boldsymbol{\pi}(s)) \ge W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))$ and $W_{\pi_j(s+1)}(\boldsymbol{\pi}(s)) < W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))$. Therefore,

$$\frac{\mu_{\pi_j(s)}}{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))} \ge \frac{\mu_{\pi_j(s)}}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))} \quad \text{and} \quad \frac{\mu_{\pi_j(s+1)}}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s))} > \frac{\mu_{\pi_j(s+1)}}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))}. \tag{37}$$

Since $\Phi(\boldsymbol{\pi}(s+1))$ differs from $\Phi(\boldsymbol{\pi}(s))$ only on the items *w.r.t.* $k = \pi_j(s)$ and $k = \pi_j(s+1)$, by examining Equation (36) and Equation (37), it is easy to check that $\Phi(\boldsymbol{\pi}(s+1))$ is greater than $\Phi(\boldsymbol{\pi}(s))$ by lexicographic order. As a result,

$$\Phi(\boldsymbol{\pi}(1)) = \Phi(\boldsymbol{\pi}(t+1)) \succ \Phi(\boldsymbol{\pi}(t)) \succ \cdots \succ \Phi(\boldsymbol{\pi}(1)), \tag{38}$$

where $\succ$ represents "greater than by lexicographic order". This leads to a contradiction. □

Proof of Condition (2). Note that when $\forall j, j' \in [N], k \in [K]$, $w_{j,k} = w_{j',k}$, we have $\forall j \in [N], k \in [K]$, $w_{j,k} = 1$ since we assume that $\forall k, \max_{j\in[N]} w_{j,k} = 1$. As a result, this condition is a special case of the condition listed in Condition (2). □

Proof of Condition (3). Suppose there exists an infinite improvement path. Then there must exist an improvement cycle

$$(\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(1)). \tag{39}$$

Let $\mathcal{K} = \{k(s) : 1 \le s \le t\}$ be the set of all resources that appear in the improvement cycle.

For each $1 \le s \le t$, since $\boldsymbol{\pi}(s+1) = \big(k(s), \boldsymbol{\pi}_{-j(s)}(s)\big)$ is an improvement step of $\boldsymbol{\pi}(s)$ and $\mu_{\pi_{j(s)}(s)} = \mu_{k(s)}$ by condition, setting $j = j(s)$ for brevity, we have

$$\frac{w_{j,\pi_j(s+1)}}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))} = \frac{w_{j,k(s)}}{w_{j,k(s)} + W_{k(s)}(\boldsymbol{\pi}(s))} > \frac{w_{j,\pi_j(s)}}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))}. \tag{40}$$

As a result, subtract the above inequality by 1 in both sides, we can get that

$$-\frac{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s))}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))} = -\frac{W_{k(s)}(\boldsymbol{\pi}(s))}{w_{j,k(s)} + W_{k(s)}(\boldsymbol{\pi}(s))} > -\frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s)) - w_{j,\pi_j(s)}}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))} = -\frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))}. \tag{41}$$

Hence

$$\frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))} > \frac{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s))}{W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))}. \tag{42}$$

As a result, we must have $W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))/W_{\pi_j(s)}(\boldsymbol{\pi}(s)) > 0$. As a result, Define

$$c_0 = \frac{1}{2} \min_{1\le s\le t} \frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))} > 0 \tag{43}$$

and a potential function $\Phi(\boldsymbol{\pi}(s))$ for any $1 \le s \le t$ as

$$\Phi(\boldsymbol{\pi}(s)) = \prod_{k\in\mathcal{K}} \left(\mathbb{I}[W_k(\boldsymbol{\pi}(s)) = 0]c_0 + \mathbb{I}[W_k(\boldsymbol{\pi}(s)) > 0]W_k(\boldsymbol{\pi}(s))\right). \tag{44}$$

By definition, we have $\Phi(s) > 0$ for any $1 \le s \le t$. Consider the change of the potential function for any round s and $s+1$. Since $W_k(\boldsymbol{\pi}(s))$ differs from $W_k(\boldsymbol{\pi}(s+1))$ only when $k = \pi_{j(s)}(s)$ and $k = \pi_{j(s)}(s+1)$. Let $j = j(s)$. By Equation (42), we have $W_{\pi_j(s)}(\boldsymbol{\pi}(s)), W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1)), W_{\pi_j(s)}(\boldsymbol{\pi}(s+1)) > 0$. Consider the two cases when $W_{\pi_j(s+1)}(\boldsymbol{\pi}(s)) = 0$ or $W_{\pi_j(s+1)}(\boldsymbol{\pi}(s)) > 0$.

- When $W_{\pi_j(s+1)}(\boldsymbol{\pi}(s)) = 0$, we have

$$\frac{\Phi(\boldsymbol{\pi}(s+1))}{\Phi(\boldsymbol{\pi}(s))} = \frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))}{W_{\pi_j(s)}(\boldsymbol{\pi}(s)) \cdot c_0} > 2 \tag{45}$$

  by the definition of c_0.

- When $W_{\pi_j(s+1)}(\boldsymbol{\pi}(s)) > 0$, we have

$$\frac{\Phi(\boldsymbol{\pi}(s+1))}{\Phi(\boldsymbol{\pi}(s))} = \frac{W_{\pi_j(s)}(\boldsymbol{\pi}(s+1))W_{\pi_j(s+1)}(\boldsymbol{\pi}(s+1))}{W_{\pi_j(s)}(\boldsymbol{\pi}(s))W_{\pi_j(s+1)}(\boldsymbol{\pi}(s))} > 1 \tag{46}$$

  by Equation (42).

As a result, we have that $\Phi(\boldsymbol{\pi}(s+1)) > \Phi(\boldsymbol{\pi}(s))$ for all $1 \le s \le t$. Therefore,

$$\Phi(\boldsymbol{\pi}(1)) = \Phi(\boldsymbol{\pi}(t+1)) > \Phi(\boldsymbol{\pi}(t)) > \cdots > \Phi(\boldsymbol{\pi}(1)), \tag{47}$$

which leads to a contradiction. Therefore, there does not exist any infinite improvement path and PNE exists. □

C.3 Proof of Theorem 4.1

Lemma C.4. *For any Nash equilibrium $\boldsymbol{\pi}$ of $\mathcal{G}$, there exists an index k, such that (1) any arm with index larger than k is not chosen by any player and (2) any arm with index not larger than k is chosen by at least one player, i.e.,*

$$\forall k' > k, \forall j \in [N], \pi_j \neq k' \quad \textit{and} \quad \forall k' \le k, \exists j \in [N], \pi_j = k'. \tag{48}$$

Proof. Suppose the condition in Equation (48) does not hold for a Nash equilibrium $\boldsymbol{\pi}$. Then there must exist two indices, $k_1 < k_2$ such that arm k_1 is not chosen by any player and arm k_2 is chosen by at least one player. Now any player that pulls arm k_2 can increase his reward by deviating to arm k_1, which leads to a contradiction since $\boldsymbol{\pi}$ is a Nash equilibrium. □

Now we can prove Theorem 4.1.

Proof of Theorem 4.1. (a) We first demonstrate that under the conditions $\Delta < \delta/(4K)$ and $\Gamma \le \delta/(4N)$, all Nash equilibria of $\mathcal{G}$ are also equilibria of $\mathcal{G}'$ and vice versa.

Note that for all $j \in N$ and $\boldsymbol{\pi} \in [K]^N$, by setting $k = \pi_j$, we have

$$\begin{aligned} \left|\mathcal{U}_j(\boldsymbol{\pi}) - \mathcal{U}'_j(\boldsymbol{\pi})\right| &= \left|\frac{(\hat{\mu}_{j,k} - \mu_k)\, w_{j,k}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = k] w_{j',k}} + \gamma_{j,k}\right| \le \left|\frac{(\hat{\mu}_{j,k} - \mu_k)\, w_{j,k}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = k] w_{j',k}}\right| + |\gamma_{j,k}| \\ &\le |\hat{\mu}_{j,k} - \mu_k| + |\gamma_{j,k}| \le \Delta + \Gamma \le \frac{\delta}{4} + \frac{\delta}{8} = \frac{3\delta}{8}. \end{aligned} \tag{49}$$

On the one hand, suppose $\boldsymbol{\pi}$ is a Nash equilibrium of $\mathcal{G}$. By the definition of δ and δ^{NE} in Equation (13), we have that

$$\forall j \in [N], \pi' \in [K] \text{ and } k \neq \pi_j, \quad \mathcal{U}_j(\boldsymbol{\pi}) - \mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) \ge \delta^{\mathrm{NE}} \ge \delta. \tag{50}$$

As a result, $\forall j \in [N], \pi' \in [K]$ and $k \neq \pi_j$, we have

$$\begin{aligned} \mathcal{U}'_j(\boldsymbol{\pi}) - \mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j}) &= \mathcal{U}_j(\boldsymbol{\pi}) - \mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) + \big(\mathcal{U}'_j(\boldsymbol{\pi}) - \mathcal{U}_j(\boldsymbol{\pi})\big) + \big(\mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j})\big) \\ &\ge \delta - \left|\mathcal{U}'_j(\boldsymbol{\pi}) - \mathcal{U}_j(\boldsymbol{\pi})\right| - \left|\mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j})\right| \ge \delta - \frac{3\delta}{8} - \frac{3\delta}{8} > 0. \end{aligned} \tag{51}$$

Therefore, $\boldsymbol{\pi}$ is also a Nash equilibrium of $\mathcal{G}'$.

On the other hand, suppose $\boldsymbol{\pi}$ is a Nash equilibrium of $\mathcal{G}'$. If $\boldsymbol{\pi}$ is not a Nash equilibrium of $\mathcal{G}$, by Equation (13), there must exist a strategy profile $\boldsymbol{\pi}$, a player j and a deviation $\pi' \neq \pi_j$ such that

$$\mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) \ge \mathcal{U}_j(\boldsymbol{\pi}) + \delta^{\mathrm{NoNE}} \ge \mathcal{U}_j(\boldsymbol{\pi}) + \delta. \tag{52}$$

As a result, we have

$$\begin{aligned}\mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}'_j(\boldsymbol{\pi}) &= \mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}_j(\boldsymbol{\pi}) + \big(\mathcal{U}_j(\boldsymbol{\pi}) - \mathcal{U}'_j(\boldsymbol{\pi})\big) + \big(\mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j})\big) \\ &\geq \delta - \big|\mathcal{U}'_j(\boldsymbol{\pi}) - \mathcal{U}_j(\boldsymbol{\pi})\big| - \big|\mathcal{U}_j(\pi', \boldsymbol{\pi}_{-j}) - \mathcal{U}'_j(\pi', \boldsymbol{\pi}_{-j})\big| \geq \delta - \frac{3\delta}{8} - \frac{3\delta}{8} > 0,\end{aligned} \tag{53}$$

which leads to a contradiction since $\boldsymbol{\pi}$ is a Nash equilibrium of $\mathcal{G}$. Therefore, $\boldsymbol{\pi}$ is also a Nash equilibrium of $\mathcal{G}$.

(b) We then demonstrate that any most efficient Nash equilibrium $\boldsymbol{\pi}$ of $\mathcal{G}'$ is also a most efficient equilibrium of $\mathcal{G}$.

Suppose $\boldsymbol{\pi}$ is a most efficient equilibrium of $\mathcal{G}'$ but not a most efficient equilibrium of $\mathcal{G}$. By the part of (a), we have that $\boldsymbol{\pi}$ is also a Nash equilibrium of $\mathcal{G}$. By Lemma C.4, we have that there exists k, such that

$$\forall k' > k, \forall j \in [N], \pi_j \neq k' \quad \text{and} \quad \forall k' \leq k, \exists j \in [N], \pi_j = k'. \tag{54}$$

And we have the total welfare $W_{\mathcal{G}'}(\boldsymbol{\pi})$ under the game $\mathcal{G}'$ is given by the following bounds:

$$\begin{aligned}W_{\mathcal{G}'}(\boldsymbol{\pi}) &= \sum_{j=1}^{N} \mathcal{U}'_j(\boldsymbol{\pi}) = \sum_{j=1}^{N} \left(\frac{\hat{\mu}_{j,\pi_j} w_{j,\pi_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = \pi_j] w_{j',\pi_j}} + \gamma_{j,\pi_j} \right) \\ &\leq \sum_{j=1}^{N} \frac{(\mu_{\pi_j} + \Delta) w_{j,\pi_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = \pi_j] w_{j',\pi_j}} + N\Gamma \\ &= \sum_{k'=1}^{k} \mu_{k'} + k\Delta + N\Gamma \leq \sum_{k'=1}^{k} \mu_{k'} + K\Delta + N\Gamma\end{aligned} \tag{55}$$

and

$$W_{\mathcal{G}'}(\boldsymbol{\pi}) = \sum_{j=1}^{N} \left(\frac{\hat{\mu}_{j,\pi_j} w_{j,\pi_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = \pi_j] w_{j',\pi_j}} + \gamma_{j,\pi_j} \right) \geq \sum_{j=1}^{N} \frac{(\mu_{\pi_j} - \Delta) w_{j,\pi_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi_{j'} = \pi_j] w_{j',\pi_j}} = \sum_{k'=1}^{k} \mu_{k'} - k\Delta \geq \sum_{k'=1}^{k} \mu_{k'} - K\Delta. \tag{56}$$

Because $\boldsymbol{\pi}$ is not a most efficient equilibrium of $\mathcal{G}$ by assumption, there must exist another equilibrium $\tilde{\boldsymbol{\pi}}$ of $\mathcal{G}$ such that $W_{\mathcal{G}}(\tilde{\boldsymbol{\pi}}) > W_{\mathcal{G}}(\boldsymbol{\pi})$. By Lemma C.4, there must exist $\tilde{k} > k$ such that in the equilibrium $\tilde{\boldsymbol{\pi}}$, all the arms in $[\tilde{k}]$ is chosen by at least one player and all other arms are not chosen by any player. By part (a), we have that $\tilde{\boldsymbol{\pi}}$ is also an equilibrium of $\mathcal{G}'$, and we have

$$W_{\mathcal{G}'}(\tilde{\boldsymbol{\pi}}) - W_{\mathcal{G}'}(\boldsymbol{\pi}) \geq \sum_{k'=1}^{\tilde{k}} \mu_{k'} - K\Delta - \left(\sum_{k'=1}^{k} \mu_{k'} + K\Delta + N\Gamma \right) \geq \mu_{\tilde{k}} - 2K\Delta - N\Gamma. \tag{57}$$

Note that we must have $\mu_{\tilde{k}} \geq \delta$ because arm $\tilde{k}$ is chosen by at least one player j in the equilibrium $\tilde{\boldsymbol{\pi}}$ *i.e.*, $\mathcal{G}$ and player j's any deviation will incur a loss not larger than $\mu_{\tilde{k}}$. As a result,

$$W_{\mathcal{G}'}(\tilde{\boldsymbol{\pi}}) - W_{\mathcal{G}'}(\boldsymbol{\pi}) \geq \mu_{\tilde{k}} - 2K\Delta - N\Gamma \geq \delta - 2K \cdot \frac{\delta}{4K} - N \cdot \frac{\delta}{4N} = \frac{\delta}{4} > 0. \tag{58}$$

This leads to a contradiction since $\boldsymbol{\pi}$ is a most efficient equilibrium of $\mathcal{G}'$. Now the claim follows.

(c) We finally demonstrate that the most efficient Nash equilibrium of $\mathcal{G}'$ is unique with probability 1.

We have

$$\begin{aligned}&\mathbb{P}\left[\exists \boldsymbol{\pi}', \boldsymbol{\pi}'' \in [K]^N, \boldsymbol{\pi}' \text{ and } \boldsymbol{\pi}'' \text{ are the most efficient Nash equilibria of } \mathcal{G}', W_{\mathcal{G}'}(\boldsymbol{\pi}') = W_{\mathcal{G}'}(\boldsymbol{\pi}'')\right] \\ \leq &\sum_{\boldsymbol{\pi}', \boldsymbol{\pi}'' \in [K]^N} \mathbb{P}[W_{\mathcal{G}'}(\boldsymbol{\pi}') = W_{\mathcal{G}'}(\boldsymbol{\pi}'')].\end{aligned} \tag{59}$$

Note that $W_{\mathcal{G}'}(\boldsymbol{\pi}') = W_{\mathcal{G}'}(\boldsymbol{\pi}'')$ if and only if

$$\sum_{j=1}^{N} \left(\gamma_{j,\pi'_j} - \gamma_{j,\pi''_j} \right) = \sum_{j=1}^{N} \left(\frac{\hat{\mu}_{j,\pi'_j} w_{j,\pi'_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi'_{j'} = \pi'_j] w_{j',\pi'_j}} - \frac{\hat{\mu}_{j,\pi''_j} w_{j,\pi''_j}}{\sum_{j'=1}^{N} \mathbb{I}[\pi''_{j'} = \pi''_j] w_{j',\pi''_j}} \right). \tag{60}$$

Note that this occurs with probability 0 because $\gamma_{j,k}$ are sampled uniformly from $[0, \Gamma]$ and the right-hand side of the above equation is a constant when $\boldsymbol{\pi}'$ and $\boldsymbol{\pi}''$ are fixed. As a result, we have $\forall \boldsymbol{\pi}', \boldsymbol{\pi}'' \in [K]^N$, $\mathbb{P}[W_{\mathcal{G}'}(\boldsymbol{\pi}') = W_{\mathcal{G}'}(\boldsymbol{\pi}'')] = 0$ and the left-hand side of Equation (59) is 0. Now the claim follows. □

C.4 Proof of Theorem 4.2

Further Notations Let $\mathcal{Z}$ be the space of all possible state of $\mathcal{M}(\epsilon; \mathcal{G}')$. In addition, define $\mathcal{D}$ as the set of all state in the Markov process where all players are discontent.

$$\mathcal{D} = \left\{ Z = \{(m_j, a_j, u_j)\}_{j=1}^N \in \mathcal{Z} : \forall j \in [N], m_j = D \right\}. \tag{61}$$

We first note that similar to [Pradelski and Young, 2012], the transition probability from any state $Z_1 \in \mathcal{Z} \backslash \mathcal{D}$ to $Z_2 \in \mathcal{D}$ and the transition probability from any state $Z_2 \in \mathcal{D}$ to $Z_1 \in \mathcal{Z} \backslash \mathcal{D}$ are independent of the Z_2. As a result, we can view all states in $\mathcal{D}$ as a single state. The transition probability from $\mathcal{D}$ to other state $Z \in \mathcal{Z} \backslash \mathcal{D}$ is the same as the probability from any state $Z_1 \in \mathcal{D}$ to Z. We slightly abuse the notation here to denote the Markov process as $\mathcal{M}(\epsilon; \mathcal{G}')$ after combining all the states in $\mathcal{D}$ as a single state.

For any $\epsilon > 0$, since the Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$ is ergodic, the stationary distribution is unique. We denote the unique stationary distribution of $\mathcal{M}(\epsilon; \mathcal{G}')$ as $p(\epsilon; \mathcal{G}')$ and $p_Z(\epsilon; \mathcal{G}')$ is the probability of state $Z \in \mathcal{Z}$ in the stationary distribution. Let $A(\epsilon; \mathcal{G}')$ be the 1/8-mixing time [Chung et al., 2012] of the Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$ and $A(\epsilon; \mathcal{G}') = \min\{t : \max_x \|xM^t - p\|_{\text{TV}} \leq 1/8\}$ where x is an arbitrary initial distribution on $\mathcal{Z}$ and $\|\cdot\|_{\text{TV}}$ denotes the total variation distance.

We first note that the process $\mathcal{M}(\epsilon; \mathcal{G}')$ is a regular perturbed Markov process defined as follows.

Definition C.1 (Regular Perturbed Markov Process [Young, 1993])**.** A Markov process $\mathcal{M}_\epsilon$ is called a regular perturbed Markov process if $\mathcal{M}_\epsilon$ is ergodic for all $\epsilon > 0$ and for every $Z, Z' \in \mathcal{Z}$ we have

$$\lim_{\epsilon \to 0^+} \mathcal{M}_\epsilon(Z \to Z') = \mathcal{M}_0(Z \to Z') \tag{62}$$

and if $\mathcal{M}_\epsilon(Z \to Z') > 0$ for some $\epsilon > 0$ then

$$0 < \lim_{\epsilon \to 0^+} \frac{\mathcal{M}_\epsilon(Z \to Z')}{\epsilon^{r(Z \to Z')}} < \infty \tag{63}$$

for some real non-negative $r(Z \to Z')$ called the resistance of the transition $Z \to Z'$.

For such regular perturbed Markov process, the stochastic stability of a state is defined as follows.

Definition C.2 (Stochastic Stability [Young, 1993])**.** A state $Z \in \mathcal{Z}$ is stochastically stable relative to the process $\mathcal{M}_\epsilon$ if

$$\lim_{\epsilon \to 0^+} p_Z(\epsilon) > 0. \tag{64}$$

Here $p_Z(\epsilon)$ denotes the probability of the state Z in the stationary distribution of $\mathcal{M}_\epsilon$.

In addition, we need the definition of interdependent game.

Definition C.3 (Interdependent Game [Pradelski and Young, 2012])**.** The game $\mathcal{G}$ is interdependent if for any strategy profile $\boldsymbol{\pi}$ and every proper subset of players $\emptyset \subsetneq J \subsetneq [N]$, there exists some player $j \notin J$ and a choice of strategies $\boldsymbol{\pi}'_J$ such that $\mathcal{U}_j(\boldsymbol{\pi}'_J, \boldsymbol{\pi}_{-J}) \neq \mathcal{U}_j(\boldsymbol{\pi})$. Here the strategy profile $(\boldsymbol{\pi}'_J, \boldsymbol{\pi}_{-J})$ denotes the profile where all players in the set J deviate to $\boldsymbol{\pi}'_J$ and other players' strategies remain unchanged.

Important Lemmas With these notations, we further need two lemmas.

Lemma C.5. *Suppose the conditions in Theorem 4.2 hold. Let $\boldsymbol{\pi}^*$ be the unique most efficient Nash equilibrium w.r.t. the game $\mathcal{G}'$ according to Theorem 4.1. Then with probability 1 the following holds. The state*

$$Z^* = \{(m_j^*, a_j^*, u_j^*)\}_{j \in [N]} \text{ with } \forall j \in [N], m_j^* = C, a_j^* = \pi_j^*, u_j^* = \mathcal{U}'_j(\boldsymbol{\pi}^*) \tag{65}$$

is the unique stochastically stable state of the regular perturbed Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$. In addition, for any $0 < \alpha < 1/2$, there exists a small enough ϵ such that the probability of Z^* in the stationary distribution $p_{Z^*}(\epsilon; \mathcal{G}')$ is larger than $1/(2(1-\alpha))$.

Proof. (a) Firstly, we demonstrate that note that $\mathcal{G}'$ is an interdependent game based on the assumption that $w_{j,k} > 0$ for all $j \in [N]$ and $k \in [K]$.

For any strategy profile $\boldsymbol{\pi}$ and every proper subset of players $\emptyset \subsetneq J \subsetneq [N]$, consider any player $j \notin J$ and any player $j' \in \mathcal{J}$. If $\pi_j = \pi_{j'}$, then the payoff of player j' will change when player j chooses another strategy instead of π_j due to $w_{j,\pi_j} > 0$. Similarly, if $\pi_j \neq \pi_{j'}$, then the payoff of player j' will also change when player j chooses the strategy $\pi_{j'}$ due to $w_{j,\pi_{j'}} > 0$. As a result, $\mathcal{G}'$ is an independent game by definition.

(b) We further demonstrate the functions $F(\cdot)$ and $G(\cdot)$ in Equation (9) meet certain conditions.

For any $j \in [N]$, define the set $V_j = \left\{\mathcal{U}'_j(\boldsymbol{\pi}) : \boldsymbol{\pi} \in [K]^N\right\}$ where $\mathcal{U}'_j(\boldsymbol{\pi})$ is provided in Equation (7). As a result,

$$\forall u' \in V_j, u' \leq \max_k(\hat{\mu}_{j,k} + \gamma_{j,k}) \leq 1 + \Gamma, \quad \text{and} \quad \forall u', u'' \in V_j, |u' - u''| \leq \max_{u' \in V_j} u' \leq 1 + \Gamma. \tag{66}$$

Therefore,

$$\forall u' \in V_j, \quad F(u') = -\frac{u'}{(4+3\Gamma)N} + \frac{1}{3N} > -\frac{1+\Gamma}{(4+3\Gamma)N} + \frac{1}{3N} > 0 \text{ and } F(u') = -\frac{u'}{(4+3\Gamma)N} + \frac{1}{3N} < \frac{1}{3N} < \frac{1}{2N}. \tag{67}$$

In addition,

$$\forall u', u'' \in V_j, u' \geq u'', \quad G(u'-u'') = -\frac{u'-u''}{4+3\Gamma} + \frac{1}{3} \geq -\frac{1+\Gamma}{4+3\Gamma} + \frac{1}{3} > 0 \text{ and } G(u'-u'') = -\frac{u'-u''}{4+3\Gamma} + \frac{1}{3} < \frac{1}{3} < \frac{1}{2}. \tag{68}$$

(c) Finally, we could prove this lemma according to Theorem 1 in [Pradelski and Young, 2012]. Note that our Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$ follow exactly the one proposed in [Pradelski and Young, 2012] and the only modifications are to ensure that the payoff calculation for each player follows the game $\mathcal{G}'$. By part (a) and (b) of our proof, the conditions of Theorem 1 in [Pradelski and Young, 2012] hold. As a result, since $\mathcal{G}'$ have at least one PNE and the most efficient PNE is unique with probability 1 according to Theorem 4.1, the only stochastically stable state is Z^* in Equation (65). As a result, when $\epsilon \to 0$, $p_{Z^*}(\epsilon; \mathcal{G}') \to 1$. Now the claim follows. □

Lemma C.6. *Suppose the conditions in Theorem 4.2 hold. Let w_s be the set of rounds in the learning phase with length $c_2 s^\eta$ and Z^* be the unique stochastically stable state of $\mathcal{M}(\epsilon; \mathcal{G}')$ given by Equation (65). Let $Z(t)$ be status of the Markov process $\mathcal{M}(\epsilon; \mathcal{G}')$ at round t. Define the counter H_s and the event E_{2s} as*

$$H_s = \sum_{t \in w_s} \mathbb{I}[Z(t) = Z^*] \quad \text{and} \quad E_{2s} = \left\{\sum_{i=1}^{s} H_i > \frac{L_s}{2}\right\}, L_s = \sum_{i=1}^{s} c_2 i^\eta. \tag{69}$$

For any $0 < \alpha < 1$, let ϵ be small enough that $p_{Z^*}(\epsilon; \mathcal{G}') > 1/(2(1-\alpha))$ according to Lemma C.5. Then there exists a constant $C(\epsilon; \mathcal{G}')$, such that

$$\mathbb{P}[E_{2s}] \geq 1 - \left(\left(1 + \frac{C(\epsilon; \mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon; \mathcal{G}')}}\right) \exp\left(-\frac{\alpha^2 c_2}{144 A(\epsilon; \mathcal{G}')}\left(p_{Z^*}(\epsilon; \mathcal{G}') - \frac{1}{2(1-\alpha)}\right)\left(\frac{s}{2}\right)^\eta\right)\right)^s, \tag{70}$$

where $A(\epsilon; \mathcal{G}')$ is the $1/8$-mixing time of $\mathcal{M}(\epsilon; \mathcal{G}')$.

Proof. The proof follows from Lemma 6 of [Bistritz and Leshem, 2021]. Define $f : \mathcal{Z} \to [0, 1]$ as $f(Z^*) = 1$ and $f(Z) = 0, \forall Z \neq Z^*$. As a result, let $\nu = \mathbb{E}_{Z \sim p(\epsilon; \mathcal{G}')}[f(Z)] = p_{Z^*}(\epsilon; \mathcal{G}')$. The initial distribution ϕ is chosen that $\phi_{\mathcal{D}} = 1$ and $\phi_Z = 0, \forall Z \neq \mathcal{D}$. According to Lemma C.7, we have

$$\mathbb{P}\left[H_s \leq (1-\alpha) p_{Z^*}(\epsilon; \mathcal{G}') c_2 s^\eta\right] \leq \frac{C(\epsilon; \mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon; \mathcal{G}')}} \exp\left(-\frac{\alpha^2 p_{Z^*}(\epsilon; \mathcal{G}') c_2 s^\eta}{72 A(\epsilon; \mathcal{G}')}\right), \tag{71}$$

Note that $\{H_s\}$ are independent. As a result, according to the Chernoff bound, we have

$$\forall \xi > 0, \quad \mathbb{P}\left[\bar{E}_{2s}\right] = \mathbb{P}\left[\sum_{i=1}^{s} H_i \leq \frac{L_s}{2}\right] \leq \exp\left(\frac{\xi L_s}{2}\right) \mathbb{E}\left[\prod_{i=1}^{s} \exp\left(-\xi H_i\right)\right] = \exp\left(\frac{\xi L_s}{2}\right) \prod_{i=1}^{s} \mathbb{E}\left[\exp\left(-\xi H_i\right)\right]. \tag{72}$$

For every i such that $1 \le i \le s$, by Equation (71) we have

$$\begin{aligned}
&\mathbb{E}[\exp(-\xi H_i)] \\
\le& \mathbb{P}\left[H_i \le (1-\alpha)p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta\right]\mathbb{E}[\exp(-\xi H_i) \mid H_i \le (1-\alpha)p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta] \\
&+ \mathbb{P}\left[H_i > (1-\alpha)p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta\right]\mathbb{E}[\exp(-\xi H_i) \mid H_i > (1-\alpha)p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta] \\
\le& \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\exp\left(-\frac{\alpha^2 p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta}{72A(\epsilon;\mathcal{G}')}\right) + \exp\left(-\xi(1-\alpha)p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta\right).
\end{aligned} \tag{73}$$

By choosing $\xi = \alpha^2/\left(72(1-\alpha)A(\epsilon;\mathcal{G}')\right)$, we have

$$\mathbb{E}[\exp(-\xi H_i)] \le \left(1 + \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\right)\exp\left(-\frac{\alpha^2 p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta}{72A(\epsilon;\mathcal{G}')}\right). \tag{74}$$

Therefore, we have

$$\begin{aligned}
\mathbb{P}[\bar{E}_{2s}] &\le \exp\left(\frac{\alpha^2 L_s}{144(1-\alpha)A(\epsilon;\mathcal{G}')}\right)\prod_{i=1}^{s}\left(1 + \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\right)\exp\left(-\frac{\alpha^2 p_{Z^*}(\epsilon;\mathcal{G}')c_2 i^\eta}{72A(\epsilon;\mathcal{G}')}\right) \\
&\le \exp\left(\frac{\alpha^2 L_s}{144(1-\alpha)A(\epsilon;\mathcal{G}')}\right)\left(1 + \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\right)^s\exp\left(-\frac{\alpha^2 p_{Z^*}(\epsilon;\mathcal{G}')L_s}{72A(\epsilon;\mathcal{G}')}\right) \\
&\le \left(1 + \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\right)^s\exp\left(-\frac{\alpha^2 L_s}{72A(\epsilon;\mathcal{G}')}\left(p_{Z^*}(\epsilon;\mathcal{G}') - \frac{1}{2(1-\alpha)}\right)\right).
\end{aligned} \tag{75}$$

Since $\forall \eta > 0$,

$$L_s = \sum_{i=1}^{s} c_2 i^\eta \ge c_2 \int_0^{s-1} i^\eta \mathrm{d}i = \frac{c_2(s-1)^{\eta+1}}{\eta+1} \ge c_2\left(\frac{s}{2}\right)^{\eta+1} \tag{76}$$

and $p_{Z^*}(\epsilon;\mathcal{G}') \ge 1/(2(1-\alpha))$, we have that

$$\mathbb{P}[\bar{E}_{2s}] \le \left(\left(1 + \frac{C(\epsilon;\mathcal{G}')}{\sqrt{p_{\mathcal{D}}(\epsilon;\mathcal{G}')}}\right)\exp\left(-\frac{\alpha^2 c_2}{144A(\epsilon;\mathcal{G}')}\left(p_{Z^*}(\epsilon;\mathcal{G}') - \frac{1}{2(1-\alpha)}\right)\left(\frac{s}{2}\right)^\eta\right)\right)^s. \tag{77}$$

Now the claim follows. □

Now we can prove Theorem 4.2.

Proof of Theorem 4.2. (a) For the exploration phase, define the clean event as

$$E_1 = \left\{\forall j \in [N], k \in [K], |\hat{\mu}_{j,k} - \mu_k| < \frac{\delta}{4K}\right\}. \tag{78}$$

Since $c_1 \ge 8K^2 \log T/\delta^2$, according to Hoeffding's inequality, we have that $\forall j \in [N], k \in [K]$,

$$\mathbb{P}\left[|\hat{\mu}_{j,k} - \mu| > \frac{\delta}{4K}\right] \le 2\exp\left(-\frac{c_1\delta^2}{8K^2}\right) \le \frac{2}{T}. \tag{79}$$

As a result, we have

$$\mathbb{P}(E_1) \ge 1 - \sum_{j=1}^{N}\sum_{k=1}^{K}\mathbb{P}\left[|\hat{\mu}_{j,k} - \mu| > \frac{\delta}{4K}\right] \ge 1 - \frac{2NK}{T}. \tag{80}$$

The expected regret cost by the bad event $\bar{E}_1$ is bounded by $(2NK/T)\cdot T\mu_{\max} = 2NK\mu_{\max} = O(1)$. As a result, we focus on the clean event for the rest of the proof.

(b) Consider the learning and exploitation phase. Under the clean event E_1, according to Theorem 4.1, we have that the most efficient Nash equilibrium $\boldsymbol{\pi}^*$ of game $\mathcal{G}'$ is also a most efficient equilibrium of game

$\mathcal{G}$. The solution is also unique with probability 1. As a result, according to Lemma C.5, for any $0 < \alpha < 1$ there exists a small enough ϵ such that the state Z^* in Equation (65) is the unique stationary distribution of $\mathcal{M}(\epsilon; \mathcal{G}')$ and $p_{Z^*}(\epsilon; \mathcal{G}') > 1/(2(1-\alpha))$.

Suppose we have switched between the learning and exploitation phase S times and the regret caused by the s-th phase is denoted as $\mathrm{Reg}_{j,s}^{\mathrm{learn}}(T)$. As a result, the expected regret in the learning and exploitation phase is given by

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid E_1\right] \\
\leq &\sum_{s=1}^{S} \mathbb{E}\left[\mathrm{Reg}_{j,s}^{\mathrm{learn}}(T) \mid E_1\right] = \sum_{s=1}^{S} \left(\mathbb{E}\left[\mathrm{Reg}_{j,s}^{\mathrm{learn}}(T) \mid E_1, E_{2s}\right] \mathbb{P}(E_{2s}) + \mathbb{E}\left[\mathrm{Reg}_{j,s}^{\mathrm{learn}}(T) \mid E_1, \bar{E}_{2s}\right] (1 - \mathbb{P}(E_{2s}))\right) \\
\leq &\sum_{s=1}^{S} \left(\mathbb{E}\left[\mathrm{Reg}_{j,s}^{\mathrm{learn}}(T) \mid E_1, E_{2s}\right] + \mu_{\max}\left(c_2 s^\eta + c_3 2^s\right) \left(B(\epsilon; \mathcal{G}') \exp\left(-\frac{\alpha^2 c_2}{144 A(\epsilon; \mathcal{G}')} \left(p_{Z^*}(\epsilon; \mathcal{G}') - \frac{1}{2(1-\alpha)}\right) \left(\frac{s}{2}\right)^\eta\right)\right)^s\right) \\
\leq &\mu_{\max} \left(\sum_{s=1}^{S} c_2 s^\eta + \sum_{s=1}^{S} \left(c_2 s^\eta + c_3 2^s\right) \left(B(\epsilon; \mathcal{G}') \exp\left(-\frac{\alpha^2 c_2}{144 A(\epsilon; \mathcal{G}')} \left(p_{Z^*}(\epsilon; \mathcal{G}') - \frac{1}{2(1-\alpha)}\right) \left(\frac{s}{2}\right)^\eta\right)\right)^s\right)
\end{aligned}
\tag{81}
$$

where $B(\epsilon; \mathcal{G}') = 1 + C(\epsilon; \mathcal{G}')/\sqrt{p_{\mathcal{D}}(\epsilon; \mathcal{G}')}$. As a result, there exists a large enough $S_0 > 0$ such that

$$
\forall s > S_0, \quad B(\epsilon; \mathcal{G}') \exp\left(-\frac{\alpha^2 c_2}{144 A(\epsilon; \mathcal{G}')} \left(p_{Z^*}(\epsilon; \mathcal{G}') - \frac{1}{2(1-\alpha)}\right) \left(\frac{s}{2}\right)^\eta\right) \leq \frac{1}{2}. \tag{82}
$$

As a result, when T is large enough (dependent on ϵ and $\mathcal{G}'$), we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid E_1\right] \\
\leq &\mu_{\max} \left(\sum_{s=1}^{S} c_2 s^\eta + \sum_{s=1}^{S_0} \left(c_2 s^\eta + c_3 2^s\right) \left(B(\epsilon; \mathcal{G}') \exp\left(-\frac{\alpha^2 c_2}{144 A(\epsilon; \mathcal{G}')} \left(p_{Z^*}(\epsilon; \mathcal{G}') - \frac{1}{2(1-\alpha)}\right) \left(\frac{s}{2}\right)^\eta\right)\right)^s\right) \\
&+ \mu_{\max} \left(\sum_{s=S_0+1}^{S} \left(c_2 s^\eta + c_3 2^s\right) \left(\frac{1}{2}\right)^s\right) \\
\leq &2\mu_{\max} \left(\sum_{s=1}^{S} c_2 s^\eta + c_3 S\right) + O(1) \\
\leq &2\mu_{\max} \left(\int_1^{S+1} c_2 s^\eta \mathrm{d}s + c_3 S\right) + O(1) \\
\leq &2\mu_{\max} \left(\frac{c_2 (S+1)^{\eta+1}}{\eta+1} + c_3 S\right) + O(1).
\end{aligned}
\tag{83}
$$

By the definition of S, we have $T \leq \sum_{s=1}^{S} c_3 2^s = c_3(2^{S+1} - 2)$ and $S = O(\log T)$. As a result, we have

$$
\mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid E_1\right] \leq 2\mu_{\max} \left(\frac{c_2 (S+1)^{\eta+1}}{\eta+1} + c_3 S\right) \leq O\left(\mu_{\max}\left(c_2 \log^{1+\eta} T + c_3 \log T\right)\right). \tag{84}
$$

(c) Combining the above two parts, when T is large, the expected regret is given by

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}_j(T)] &= \mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{init}}(T)\right] + \mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T)\right] \\
&\leq \mu_{\max} c_1 + \mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid E_1\right] \mathbb{P}(E_1) + \mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid \bar{E}_1\right] (1 - \mathbb{P}(E_1)) \\
&\leq \mu_{\max} c_1 + \mathbb{E}\left[\mathrm{Reg}_j^{\mathrm{learn}}(T) \mid E_1\right] + 2NK\mu_{\max} \\
&\leq \mu_{\max} c_1 + O\left(\mu_{\max}\left(c_2 \log^{1+\eta} T + c_3 \log T\right)\right) \\
&= O\left(\mu_{\max}\left(c_1 + c_2 \log^{1+\eta} T + c_3 \log T\right)\right).
\end{aligned}
\tag{85}
$$

Similar to the above proof, we have that

$$\begin{aligned}
\mathbb{E}[\mathrm{NonPNE}(T)] &= \mathbb{E}\left[\mathrm{NonPNE}^{\mathrm{init}}(T)\right] + \mathbb{E}\left[\mathrm{NonPNE}^{\mathrm{learn}}(T)\right] \\
&\leq c_1 + \mathbb{E}\left[\mathrm{NonPNE}^{\mathrm{learn}}(T) \mid E_1\right] \mathbb{P}(E_1) + \mathbb{E}\left[\mathrm{NonPNE}^{\mathrm{learn}}(T) \mid \bar{E}_1\right] (1 - \mathbb{P}(E_1)) \\
&\leq c_1 + \mathbb{E}\left[\mathrm{NonPNE}^{\mathrm{learn}}(T) \mid E_1\right] + 2NK \\
&\leq c_1 + O\left(c_2 \log^{1+\eta} T + c_3 \log T\right) \\
&= O\left(c_1 + c_2 \log^{1+\eta} T + c_3 \log T\right).
\end{aligned} \tag{86}$$

Here $\mathrm{NonPNE}^{\mathrm{init}}(T)$ and $\mathrm{NonPNE}^{\mathrm{learn}}(T)$ represents the number of rounds where players do not follow a most efficient PNE in the exploration phase and learning and exploitation phase, respectively. The bounds on $\mathrm{NonPNE}^{\mathrm{init}}(T)$ and $\mathrm{NonPNE}^{\mathrm{learn}}(T)$ share the same techniques with $\mathrm{Reg}_j^{\mathrm{init}}(T)$ and $\mathrm{Reg}_j^{\mathrm{learn}}(T)$. Now the claim follows. □

C.5 Important Lemmas

Lemma C.7 (Theorem 3.1 in [Chung et al., 2012])**.** *Let M be an ergodic Markov chain with state space $\mathcal{Z}$ and stationary distribution p. Let A be its $1/8$-mixing time. Let $(Z_1, \ldots, Z_t)$ denote a t-step random walk on M starting from an initial distribution ϕ on $\mathcal{Z}$, i.e., $Z_1 \sim \phi$. For every $i \in [t]$, let $f_i : \mathcal{Z} \to [0, 1]$ be a weight function at step i such that the expected weight $\mathbb{E}_{Z\sim p}[f_i(Z)] = \nu$ for all i. Define the total weight of the walk $(Z_1, \ldots, Z_t)$ by $Y = \sum_{i=1}^{t} f_i(Z_i)$. There exists some constant c (which is independent of ν and α) such that $\forall 0 < \alpha < 1$,*

$$\mathbb{P}[Y \leq (1 - \alpha)\nu t] \leq c\|\phi\|_p \exp\left(-\alpha^2 \nu t/(72A)\right). \tag{87}$$

Here $A = \min\{t : \max_x \|xM^t - p\|_{TV} \leq 1/8\}$ where x is an arbitrary initial distribution on $\mathcal{Z}$ and $\|\cdot\|_{TV}$ denotes the total variation distance. In addition, $\|\phi\|_p = \sqrt{\sum_{Z\in\mathcal{Z}} \phi^2(Z)/p(Z)}$.