Faster Convergence for Unknown-Game Banditsx

Zhiming Huang, Jianping Pan
Department of Computer Science, University of Victoria, BC, Canada

***Abstract*—In this paper, we study unknown-game bandits, where multiple agents play a general-sum game repeated over T rounds. In each round, each agent independently selects an action and observes the reward for that action. The game is *unknown* to every agent, meaning each agent has no knowledge about the underlying game structure, the number of other agents, or their actions and rewards. Such unknown-game bandits have wide applications in computer and communication networks, including congestion control and network selection. The goal of each agent is to minimize *swap regret*, which measures the performance gap from a broader class of competitors than the traditional *external regret* that only compares against competitors always playing a fixed action. Our main contribution is to bridge the gap in the literature by proving the first swap-regret bound with a time-dependence of $\tilde{O}(T^{\frac{1}{4}})$ if the proposed learning algorithm based on *optimistic follow-the-regularized-leader (OFTRL)* is played by all agents involved in the game, where $\tilde{O}(\cdot)$ hides logarithmic factors. This regret bound demonstrates a faster convergence rate with respect to the number of rounds T compared to the state-of-the-art swap regret bound of $O(T^{\frac{1}{2}})$. Furthermore, we demonstrate the efficacy of the proposed algorithm through an application in heterogeneous network selection with both numerical and simulation-based experiments.**

I. Introduction

Multi-armed bandits are online learning models designed to balance exploration and exploitation in sequential decision-making problems. In the basic bandit model, an agent selects an arm (i.e., an action) in each round and observes the reward only for the chosen arm. The objective is to minimize the *external regret* [1], [2], which compares the performance between the agent and a class of competitors always playing a fixed action.

In this paper, we study unknown-game bandits [1], [3]–[6], which capture the essence of many problems in computer and communication networks, such as congestion control [7]–[9] and heterogeneous network selection [10]–[12]. In end-to-end congestion control, each (TCP) flow must decide the sending rate to the network, treating the network as a black box. The only available knowledge is the feedback (e.g., throughput, RTT, jitters, etc.) observed from prior decisions. Similarly, in the heterogeneous wireless network selection problem, each client must select a network to attach to and can only observe the feedback (e.g., throughput, packet loss rate, etc.) from the attached network.

A commonality in these problems is that multiple agents independently make *uncoupled* decisions, each striving to optimize its own objectives in competition for limited resources. In end-to-end congestion control, if multiple agents send more packets than the network can handle, congestion occurs, leading to packet loss. Similarly, in heterogeneous network selection, if multiple devices attach to the same network, the throughput for all those devices is adversely affected. The term "uncoupled" indicates that each agent interacts independently with a black-box environment, basing its strategy solely on feedback from its own actions. Due to protocol constraints or privacy considerations, direct communication between agents is restricted, necessitating independent optimization with limited information.

Formally, we consider unknown general-sum games, or black-box games as studied in [13], where a set of agents are playing an unknown general-sum game repeated for T rounds. In each round, each agent needs to play an action and observes the corresponding reward. The game is *unknown*, because each agent is unaware of the underlying game structure, the number of other agents, or their actions. The only information available to each agent is the observed reward from their own actions.

The objective for each agent is to minimize *swap regret* [4]. Swap regret is a stronger notion than external regret, as it compares against a broader class of competitors that include those considered by external regret. Minimizing swap regret offers two significant advantages. First, a no-swap-regret algorithm demonstrates robustness against a wider range of competitors compared to a no-external-regret algorithm, where *no regret* implies that time-averaged regret vanishes as time approaches infinity. Second, minimizing swap regret can lead to convergence to a set of *correlated equilibria (CE)* [14], a concept more general than the well-known Nash equilibrium. CE ensures that no agent finds it beneficial to unilaterally change their strategy given the joint distribution of all agents' actions, thereby achieving optimal overall performance for all agents.

In network applications, the slow convergence rate often poses a challenge for learning algorithms. However, unknown-game bandits offer a potential solution with faster convergence rates. For these bandits, the sum of external regret across all agents can scale with time as $\tilde{O}(T^{\frac{1}{4}})$ [15], implying a bound of $\tilde{O}(T^{-\frac{3}{4}})$ on the sum of *time-averaged* external regret. In contrast, in scenarios with full-information feedback (where rewards for all actions are observable), the time-dependence of swap regret can be significantly reduced to $O(\ln T)$. However, for bandit feedback, the best-known swap-regret bound has a time-dependence of $O(T^{\frac{1}{2}})$ [5], [6], [16]. This raises a fundamental question: can we achieve a faster convergence rate of swap regret for unknown-game bandits?

Our work makes a significant contribution by narrowing the gap in the literature, achieving a $\tilde{O}(T^{\frac{1}{4}})$ time-dependence for swap regret. This faster convergence is realized through the adoption of OFTRL-LogBar-Bandit algorithms by all agents.

Our approach refines the BM-OFTRL-LogBar method from [17] to accommodate bandit feedback. This refinement is non-trivial; it involves careful design of the reward estimator and prediction vectors to ensure the convergence of the OFTRL algorithm while accelerating the convergence rate. Further details can be found in Sections IV and V. Our algorithm builds upon the swap-regret-minimizing framework proposed by [4], utilizing A_n OFTRL subroutines with logarithmic barrier regularizers, where A_n denotes the number of actions available to agent n. Additionally, we demonstrate in Sec. VI the efficacy of our approach in the context of heterogeneous network selection with both numerical and simulation-based experiments.

II. Related Works

Learning for Unknown Games: The pursuit of strategies for unknown games traces its roots back to the concept of fictitious play in two-agent zero-sum games [18], [19], which relies on knowledge of opponents' past plays. However, the effective tackling of challenges posed by unknown games only became feasible with the advent of online learning techniques. This development uncovered an intrinsic connection between game equilibria and strategies for regret minimization. Specifically, minimizing external regret can lead to Nash equilibria in two-agent zero-sum games and coarse correlated equilibria in multi-agent general-sum games [1]. Furthermore, minimizing swap regret, a stronger notion than external regret, has been shown to lead to correlated equilibria in multi-agent general-sum games [4], [20], [21].

When considering the observability of rewards, two distinct research lines emerge: *full-information feedback* and *bandit feedback*. While steady progress has been made in learning general-sum games under the full-information model [22]–[27], extending these results to bandit feedback poses challenges due to the limited information available in each round. The earliest work on bandit feedback dates back to [3], which introduced an exponential-weighting-based technique for minimizing external regret.

To achieve the correlated equilibrium, the authors of [28] proposed an algorithm but with an exponential computation complexity. Subsequently, the authors of [4] introduced the stronger notion of swap regret and devised a framework to efficiently transform external-regret-minimizing algorithms into swap-regret-minimizing ones with a polynomial computation complexity. Building on this framework, subsequent research has aimed to improve the swap regret bounds for bandit feedback [5], [6], [16], resulting in the recognition that the best swap regret bound for bandit feedback now achieves a time-dependence of $O(T^{\frac{1}{2}})$.

Recently, more efficient conversion techniques with faster convergence rates have been introduced for scenarios where a distribution over actions is played directly in each round, rather than sampling and playing a single action from the distribution [20], [21]. In this paper, we still focus on playing a single action in each round. Although it is known that $O(T^{\frac{1}{2}})$ is minimax-optimal in adversarial environments, there has been progress, as discussed below, that faster regret convergence is possible for learning agents in the unknown-game environment.

Faster Regret Convergence for Unknown Games: In this paper, regret convergence refers to the vanishing of time-averaged regret over time, which is different from the notions of last-iterate convergence [29], [30] and frequent-iterate convergence [31] in literature.

While faster regret convergence for bandits remains an ongoing area of exploration, significant advancements have been made in understanding unknown games with full-information feedback over the last decade. The seminal work by [32] demonstrated that the sum of external regret for all agents can be bounded by $O(1)$, with individual external regret scaling as $O(T^{\frac{1}{4}})$. Building on this foundation, the authors of [24] improved the individual external regret bound to $O(T^{\frac{1}{6}})$ for two-agent games and extended these results to swap regret, achieving a time-dependence of $O(T^{\frac{1}{4}})$.

Further progress was made by [25], who demonstrated that $O(\ln T)$ dependence for individual external regret is achievable in multi-agent general-sum games. Recently, the authors of [17], [26] achieved breakthroughs in swap regret, showing that a time-dependence of $O(\ln T)$ can be attained.

However, all the above progress is for full-information feedback. Regarding the bandit feedback, there is still a large gap in the literature. The only results are from [15], where they proved the sum of external regret for two-agent zero-sum games enjoys a time-dependence of $O(T^{\frac{1}{4}})$. Neither the individual external regret nor the swap regret is guaranteed. Thus, it remains an open question in the literature whether the time-dependence of $\tilde{O}(T^{\frac{1}{2}})$ for swap regret can be further improved.

In this paper, we take a step forward to this open question by showing a swap regret with time-dependence of $\tilde{O}(T^{\frac{1}{4}})$. Given that swap regret is a stronger measure than external regret and is always non-negative, our results also guarantee that individual external regret exhibits a time-dependence of $\tilde{O}(T^{\frac{1}{4}})$.

III. Problem Formulation

We consider general-sum games involving N agents, repeated over T rounds. Each agent $n \in [N] := \{1, \dots, N\}$ has an action set $\mathcal{A}_n$ of finite size $A_n := |\mathcal{A}_n|$ and a reward function $u_n : \mathcal{A} \to [0, 1]$, which maps joint actions of all agents $\mathcal{A} := \bigotimes_{n=1}^{N} \mathcal{A}_n$ to values in the range $[0, 1]$.

In each round t, each agent $n \in [N]$ plays an action $a_n^t \in \mathcal{A}_n$ according to a mixed strategy $p_n^t \in \Delta(\mathcal{A}_n) := \{p \in \mathbb{R}_{\geq 0}^{A_n} : \sum_{a \in \mathcal{A}_n} p(a) = 1\}$, i.e., a probability distribution among action set $\mathcal{A}_n$. The agent then observes the reward for the played action, $u_n^t(a_n^t) := \mathbf{E}_{A_{-n}^t \sim p_{-n}^t}[u_n(a_n^t; A_{-n}^t)]$, where A_{-n}^t are the actions chosen by agents other than n, and p_{-n}^t are their corresponding mixed strategies.

Note that all agents operate in a black-box setting with bandit feedback, meaning they have limited knowledge about the environment, such as the underlying game structure, the number of agents, or the actions of other agents. The only

information available to them is the observed rewards for their own played actions in each round. Furthermore, the reward for each agent depends on the joint actions of all agents in each round, creating a competitive environment. Such a general game setting is applicable to numerous critical network problems, including end-to-end congestion control and heterogeneous network selection.

In fact, each agent in unknown-game bandits faces a special multi-armed bandit problem. In single-agent bandit problems, the primary objective is to design a learning algorithm to minimize *external regret*, which is the performance gap compared to competitors always playing a fixed action. However, in unknown-game bandits, where multiple agents act adaptively to compete with each other, minimizing external regret alone does not ensure optimal overall performance for all agents. Instead, each agent $n \in [N]$ aims to compare with a broader class of competitors $\mathcal{F}_n := \{F : \mathcal{A}_n \to \mathcal{A}_n\}$, which results in the following definition of *swap regret*:

$$R_n^{\mathrm{swa}}(T) = \max_{F \in \mathcal{F}_n} \mathbf{E}\left[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}_n} \mathbf{1}[a_n^t = a]\left(u_n^t(F(a)) - u_n^t(a)\right)\right]. \tag{1}$$

We can reduce swap regret to external regret by restricting $\mathcal{F}_n$ to the subset $\{F_a, \forall a \in \mathcal{A}_n : \mathcal{A}_n \to a\}$. Therefore, swap regret is a stricter notion than external regret, and minimizing swap regret will also minimize external regret.

More importantly, if all agents employ a learning algorithm that can minimize the swap regret, their expected empirical joint distribution of plays, i.e., $\mathbf{E}[\frac{1}{T}\sum_{t=1}^{T}\sum_{A \in \mathcal{A}} \mathbf{1}[A_t = A]]$ converges to the set of *ϵ-correlated equilibria (CE)* [1], [4], where $A^t := (a_1^t, \ldots, a_N^t)$ is the joint actions played by all agents in round t. The notion of ϵ-CE is defined as follows.

Definition III.1 (ϵ-Correlated Equilibrium [14])**.** A joint probability distribution $\mathbf{P}$ over $\mathcal{A}$ is an ϵ-CE if the expected incentive for each agent n to deviate from action a to any other action $a' \in \mathcal{A}_n$ is no more than $\epsilon \geq 0$, i.e., $\forall n \in \mathcal{N}$, we have

$$\sum_{(a;A_{-n}) \in \mathcal{A}} \mathbf{P}((a;A_{-n}))\left(u_n(a';A_{-n}) - u_n(a;A_{-n})\right) \leq \epsilon. \tag{2}$$

CE is a more general concept than the well-known *Nash equilibrium (NE)*. Intuitively, CE states that if there is a mediator sampling a joint action from $\mathbf{P}$ and privately signaling the action to each agent, no agent would deviate from the mediator's suggestion. Consequently, achieving CE typically results in better overall performance compared to achieving NE.

Thus, swap regret is an important performance metric in unknown-game bandits. It not only extends the concept of external regret, used as a performance metric in standard bandits, but also guarantees convergence to ϵ-CE. The primary objective of this work is to achieve a faster convergence rate for swap regret, i.e., minimizing the dependence of $R_n^{\mathrm{swa}}(T)$ on T as much as possible.

IV. The OFTRL-LogBar-Bandit Algorithm

The OFTRL-LogBar-Bandit algorithm is designed for independent execution by all agents to achieve a faster convergence rate, as described in Alg. 1. It refines the full-information algorithm BM-OFTRL-LogBar [17], incorporating new prediction vectors and reward estimators tailored for bandit feedback [33], [34].

Algorithm 1 The OFTRL-LogBar-Bandit algorithm

Input: $n, \mathcal{A}_n, \eta$
// Initialization
Set $q_a^t(a') = \frac{1}{A_n}, \forall a, a' \in \mathcal{A}_n$
for $t = 1, \ldots, T$ **do**
  // Compute the sample distribution, play arms and observe rewards
  Calculate p_n^t based on (3)
  Play an action $a_n^t \sim p_n^t$
  Construct the estimated reward $\hat{u}_n^t$ according to (5)
  // Update each meta-distribution
  for $a \in \mathcal{A}_n$ **do**
    Update q_a^{t+1} according to (4)
  end for
end for

The main idea of OFTRL-LogBar-Bandit is to use the swap-regret-minimizing framework introduced by [4], calling A_n *optimistic follow-the-regularized-leader (OFTRL)* algorithms with the log-barrier regularizer [17] as subroutines. Since OFTRL-LogBar-Bandit runs independently for each agent, we will fix an agent $n \in [N]$ and describe how OFTRL-LogBar-Bandit operates.

Each subroutine is indexed by $a \in \mathcal{A}_n$, and maintains a meta-distribution $q_a^t \in \Delta(\mathcal{A}_n)$ by following the OFTRL framework with log-barrier regularizer. Then, let Q_n^t be a $A_n \times A_n$ stochastic matrix with each row being $(q_a^t)^\top$. The action selection probability p_n^t is then calculated from the meta-distributions as follows:

$$(p_n^t)^\top = (p_n^t)^\top Q_n^t, \tag{3}$$

which is equivalent to calculating the stationary distribution of a Markov chain described by Q_n^t.

Next, we give details on how each subroutine maintains q_a^t. Denote by $\hat{u}_n^t$ the estimated reward vector observed by OFTRL-LogBar-Bandit for all actions in round t, where the construction of $\hat{u}_n^t$ will be introduced later. Then, each subroutine $a \in \mathcal{A}_n$ observes a portion of the estimated reward by $p_n^t(a)\hat{u}_n^t$ and calculates q_a^t by solving the following optimization problem:

$$q_a^t := \operatorname*{arg\,max}_{q \in \Delta(\mathcal{A}_n)} \left\{ \eta \left\langle q, p_n^{t-1}(a) m_n^t + \sum_{s=1}^{t-1} p_n^s \hat{u}_n^s \right\rangle + \sum_{a' \in \mathcal{A}_n} \ln(q(a')) \right\}, \tag{4}$$

where m_n^t is the predictor vector to make FTRL "optimistic". To see this, if one were able to obtain $\hat{u}_n^t$ in advance and let $m_n^t := \hat{u}_n^t$, the best q_a^t can be found. Thus, if one can construct m_n^t closer to $\hat{u}_n^t$, better performance can be achieved.

In this work, we follow the convention in [15] to construct m_n^t as follows but with a key challenge as discussed at the end of this section. Let $\tau_t(a) := \arg\max_{s<t}\{\mathbf{1}[a_n^s = a]\}$ denote the last round before t when action a was played. Then, let $m_n^t(a) := u_n^{\tau_t(a)}(a)$ for all $a \in \mathcal{A}_n$, i.e., we set the predictor for each action a to be its last observed reward. For analytical convenience, and without loss of generality, we set $u_n^0(a) = 0$ and $p_n^0(a) = \frac{1}{A_n}$ for all $a \in \mathcal{A}_n$.

Since we consider bandit feedback, only the reward for the played action can be observed. Therefore, we need to construct a reward estimator that, in expectation, is equivalent to the reward received in the full-information setting:

$$\hat{u}_n^t(a') = m_n^t(a') - \frac{(m_n^t(a') - u_n^t(a'))\mathbf{1}[a_n^t = a']}{p_n^t(a')}. \quad (5)$$

A key distinction between our reward estimator and that used in [15] is that each subroutine in our algorithm receives only a portion of $\hat{u}_n^t(a')$. In [15], the entire reward estimator is fed into the algorithm, allowing m_n^t in the algorithm to be canceled out by its counterpart in the reward estimator, which simplifies regret analysis. However, such cancellation does not occur in swap-regret analysis, where an additional term $(p_n^t(a) - p_n^{t-1}(a))m_n^t$ introduces further analytical complexity.

V. Analytical Results for OFTRL-LogBar-Bandit

We begin by introducing the notations specific to our regret analysis. Denote by $\|x\|_{q_a^t} := \sqrt{\sum_{a' \in \mathcal{A}_n} (\frac{x(a')}{q_a^t(a')})^2}$ the primal local norm of vector $x \in \mathbb{R}^{A_n}$ with regard to q_a^t, and by $\|x\|_{*,q_a^t} := \sqrt{\sum_{a' \in \mathcal{A}_n} (x(a')q_a^t(a'))^2}$ the dual local norm. Furthermore, denote by $\mathcal{F}_t$ the σ-algebra formed by the history of all agents' plays and rewards up to the end of round t.

A. Regret Bounds

The following lemma is one of our key contributions that play a crucial role in deriving a faster convergence rate for regret bounds, which bounds the gap between the predictor m_n^t compared with the estimated rewards $\hat{u}_n^t$ for all subroutines over T rounds.

Lemma V.1. *For any $n \in [N]$, we have that*

$$\sum_{t=1}^{T} \sum_{a \in \mathcal{A}_n} \left\| p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t \right\|_{*,q_a^t}^2 \le 2\sum_{t=1}^{T} \|u_n^t - u_n^{t-1}\|_1 + 2\sum_{t=1}^{T} \|p_n^t - p_n^{t-1}\|_1.$$

Proof Sketch. The proof follows two steps, and the details for each step can be found in the Appendix B.

1) By applying Cauchy-Schwarz inequality and the definition of the dual local norm, we first prove that for any $t \in [T]$ and $n \in [N]$:

$$\sum_{a \in \mathcal{A}_n} \left\| p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t \right\|_{*,q_a^t}^2 \le 2\sum_{a \in \mathcal{A}_n} \left| u_n^t(a) - m_n^t(a) \right| \mathbf{1}[a_n^t = a] + 2\left\| p_n^t - p_n^{t-1} \right\|_1. \quad (6)$$

2) The rest is to bound the first term when summed over T rounds. Recall the $m_n^t(a) = u_n^{\tau_t(a)}(a)$ is the last-round reward for arm a before t. We can rewrite it as a telescoping series and apply the triangle inequality to bound it:

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{a \in \mathcal{A}_n} \left| u_n^t(a) - m_n^t(a) \right| \mathbf{1}[a_n^t = a] \\
&\le \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_n} \mathbf{1}[a_n^t = a] \sum_{s=\tau_t(a)+1}^{t} |u_n^s(a) - u_n^{s-1}(a)| \\
&= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_n} |u_n^t(a) - u_n^{t-1}(a)|,
\end{aligned}$$

where the last equality is due to $\mathbf{1}[a_n^t = a]$ and that $\tau_t(a)$ is the last time before t when a is played.

□

The above lemma converts the gap between estimated rewards and predictors for each subroutine into the gap between actual observed rewards and the difference between mixed strategies in two consecutive rounds. This helps derive the following individual swap regret for each agent.

Theorem V.2. *For $\eta \le \frac{1}{32}$, the swap regret for each agent $n \in [N]$ is upper bounded as follows.*

$$R_n^{\mathrm{swa}}(T) \le \frac{2(A_n)^2 \ln T}{\eta} + 2\eta \sum_{t=1}^{T} \sum_{m \in [N]} \left\| p_m^t - p_m^{t-1} \right\|_1 - \frac{1}{1024 A_n \eta} \sum_{t=1}^{T} \left\| p_n^t - p_n^{t-1} \right\|_1^2.$$

Proof Sketch. The proof follows the four steps. The details can be found in the Appendix D.

1) We first convert the regret defined in (1) as follows:

$$R_n^{\mathrm{swa}}(T) = \max_{F \in \mathcal{F}_n} \sum_{a \in \mathcal{A}_n} \mathbf{E}\left[\underbrace{\sum_{t=1}^{T} p_n^t(a)\hat{u}_n^t(F(a)) - \left\langle q_a^t, p_n^t(a)\hat{u}_n^t \right\rangle}_{=: R_a^T} \right].$$

Such a conversion requires the application of the tower rule on $\sum_{a \in \mathcal{A}_n} p_n^t(a)\hat{u}_n^t$ and the definition of p_n^t in (3).

2) Next, notice that R_a^t can be bounded as follows:

$$R_a^T \le \max_{q \in \Delta(\mathcal{A}_n)} \sum_{t=1}^{T} \left\langle q - q_a^t, p_n^t(a)\hat{u}_n^t \right\rangle,$$

and the RHS of the above equation can be bounded by invoking Corollary B.3 of [17] as follows:

$$R_a^T \le \frac{2A_n \ln T}{\eta} + 2\eta \sum_{t=1}^{T} \left\| p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t \right\|_{*,q_a^t}^2 - \frac{1}{16\eta} \sum_{t=1}^{T} \left\| q_a^t - q_a^{t-1} \right\|_{q_a^{t-1}}^2. \quad (7)$$

While leveraging the results from Corollary B.3 of [17], our proof is non-trivial due to several analytical challenges. First, our prediction vector m_n^t differs from those used in full-information settings, necessitating a meticulous analysis to ensure that the conditions for applying Corollary B.3 still hold with this adapted vector for bandit feedback. Second, Corollary B.3 provides a standard analysis for OFTRL with self-concordant functions; our primary challenge lies in bounding the last two terms on the right-hand side of (7), as discussed in Lemmas V.1 and A.3. These bounds require careful consideration, particularly since m_n^t lacks an indicator function typically found in reward estimators.

3) The swap regret is bounded by summing over $a \in \mathcal{A}_n$, using Lemma V.1 to bound the second term on the RHS of the above inequality, and applying Lemma A.3 (see Appendix C) to bound that $\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_n}\left\|q_a^t - q_a^{t-1}\right\|^2_{q_a^{t-1}} \geq \frac{1}{64A_n}\sum_{t=1}^{T}\left\|p_n^t - p_n^{t-1}\right\|_1^2$. Then, we obtain

$$\sum_{a\in\mathcal{A}_n} R_a^T \leq \frac{2(A_n)^2\ln T}{\eta} + 2\eta\sum_{t=1}^{T}\|u_n^t - u_n^{t-1}\|_1 + 2\eta\sum_{t=1}^{T}\|p_n^t - p_n^{t-1}\|_1 - \frac{1}{1024A_n\eta}\sum_{t=1}^{T}\left\|p_n^t - p_n^{t-1}\right\|_1^2.$$

Theorem V.2 follows by proving $\|u_n^t - u_n^{t-1}\|_1 \leq \sum_{m\in[N]\setminus n}\left\|p_m^t - p_m^{t-1}\right\|_1$, which utilizes the facts that u_n^t is determined by all agents' joint actions, and the total distance between two product distributions is bounded by the sum of the total variations of each marginal distribution [35].

□

We are now prepared to prove our main claim regarding the faster convergence rate. Let $A_{\max} := \max_{n\in[N]} A_n$. By summing the individual swap regret for all agents $n \in [N]$, we arrive at the following corollary.

Corollary V.3. *When $\eta = \frac{1}{32}(\ln T/T)^{\frac{1}{4}}N^{-\frac{1}{2}}$, we have*

$$\sum_{n\in[N]} R_n^{\mathrm{swa}}(T) \leq 65N^{\frac{3}{2}}(A_{\max})^2T^{\frac{1}{4}}(\ln T)^{\frac{3}{4}}.$$

Proof. Since $\eta \leq \frac{1}{32}$, invoking Theorem V.2 gives

$$\sum_{n\in[N]} R_n^{\mathrm{swa}}(T) \leq \frac{2N(A_{\max})^2\ln T}{\eta} + 2\eta N\sum_{t=1}^{T}\sum_{n\in[N]}\left\|p_n^t - p_n^{t-1}\right\|_1 - \frac{1}{1024A_{\max}\eta}\sum_{n\in[N]}\sum_{t=1}^{T}\left\|p_n^t - p_n^{t-1}\right\|_1^2.$$

Let $x_n^t := \left\|p_n^t - p_n^{t-1}\right\|_1$, and we have that

$$\begin{aligned}\sum_{n\in[N]} R_n^{\mathrm{swa}}(T) &\leq \frac{2N(A_{\max})^2\ln T}{\eta} + 1024\eta^3A_{\max}N^3T \\ &\quad - \frac{1}{1024\eta A_{\max}}\sum_{t=1}^{T}\sum_{n\in[N]}\left(x_n^t - 1024\eta^2A_{\max}N\right)^2 \\ &\leq \frac{2N(A_{\max})^2\ln T}{\eta} + 1024\eta^3A_{\max}N^3T.\end{aligned}$$

The corollary follows by substituting $\eta = \frac{1}{32}(\ln T/T)^{\frac{1}{4}}N^{-\frac{1}{2}}$ into the above inequality. □

We have established that the sum of the swap regret for all N agents is $\tilde{O}\left(N^{\frac{3}{2}}(A_{\max})^2T^{\frac{1}{4}}\right)$, where $\tilde{O}(\cdot)$ hides the logarithmic factors. This implies that the time-averaged swap regret for all agents is $\tilde{O}\left(N^{\frac{3}{2}}(A_{\max})^2T^{-\frac{3}{4}}\right)$, indicating that the swap regret decays at a rate of $\tilde{O}(T^{\frac{3}{4}})$. This is a significant improvement over previous results for bandit [4]–[6], [16] which depend on $O(\sqrt{T})$ for their swap regret bounds, i.e., the time-averaged regret decays at a rate of $O(T^{\frac{1}{2}})$. The intuition behind the improvement lies in leveraging the fact that each agent employs a swap-regret-minimizing algorithm. This behavior, being predictable through the OFTRL framework, facilitates a convergence speed-up.

Since swap regret is always non-negative for each agent $n \in [N]$, Corollary V.3 also implies the individual swap regret decays at a rate at least of $\tilde{O}(T^{\frac{3}{4}})$. Because bounding swap regret also bounds external regret, this provides a stronger guarantee than the results in [15]. While they demonstrated a faster convergence rate for the sum of external regret, their results do not guarantee the same rate for the individual external regret, as external regret for some agents can be negative [36].

Compared with the results for the full-information setting [17], [26], which demonstrate a $\tilde{O}(T^{-1})$ time-dependence for time-averaged swap regret, smaller than our time-dependence of $\tilde{O}(T^{-\frac{3}{4}})$. It remains an open problem whether the convergence rate for the bandit feedback can be further improved.

B. Time and Space Complexity

The time and space complexities are similar to those in previous works [6], [16] based on the swap-regret-minimizing framework [4]. Note that the optimization problem in (4) is a linear optimization with self-concordant functions, which has efficient solutions, making the primary computational complexity stem from calculating the stationary distribution of the Markov chain with A_n states. This stationary distribution can be precisely computed in $O(A_n^2)$ time [37], and approximately computed in almost linear time [38]. Regarding the space complexity, it is $O(A_n^2)$ because each subroutine process needs to maintain meta-distributions and reward vectors for A_n actions. With no communications between agents, there is no communication overhead.

VI. Experiments

Heterogeneous network selection is a typical application for unknown-game bandits [10], [11]. In this context, each device acts as an agent in the unknown-game bandit model, deciding which heterogeneous network—such as LTE, WiFi, and 5G—to attach to. Therefore, the action set for each agent consists of the available heterogeneous networks, and the reward corresponds to the observed PHY rates of the network to which the agent attaches.

In this section, we adopted experiment settings consistent with those in [10] for both numerical and simulation experiments. Our aim is to compare our OFTPL-based algorithm with two online learning algorithms—Lights and RLNF—studied in [10], [11], showing our faster regret convergence. Lights [10] is also based on the swap-regret minimizing framework [4] and employs the exponentially weighted technique [3]. On the other hand, RLNF is based on a regret minimization approach proposed in [39].

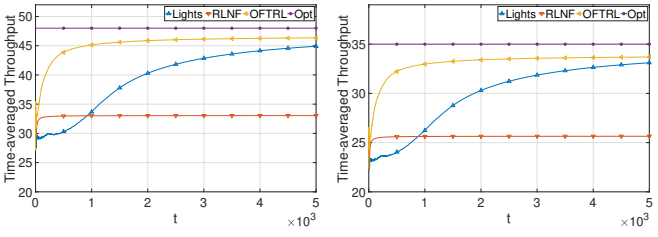
The experiment settings are described as follows.

- Setting 1 is a numerical experiment, focusing on a game between two clients connecting to LTE and WiFi networks. The game is characterized by the reward matrix defined in Table I, which specifies the maximum throughput in Mbps achievable for each client.
- Setting 2 utilizes the Matlab Communications Toolbox™ Wireless Network Simulation Library. This scenario involves 20 clients following a waypoint mobility model within a square area measuring 150 meters by 150 meters. Base stations equipped with the latest technologies, including IEEE 802.11be WiFi and 5G, are strategically deployed throughout the area (refer to Fig. 2 in [10] for further details, which are used for comparison).

Table I: The Unnormalized Reward Matrix for Setting 1 [10].

		Client 2	
		LTE	WiFi
Client 1	LTE	(17.5,17.5)	(35,24)
	WiFi	(48,35)	(16,16)

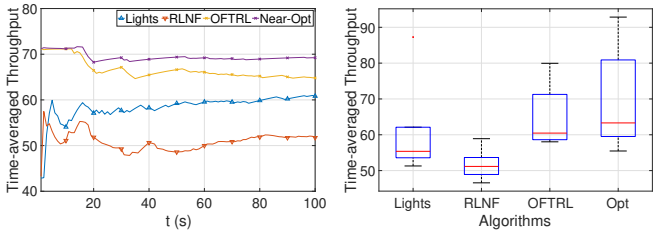
A. Time-averaged Throughput



(a) Client 1 (b) Client 2

Figure 1: The time-averaged throughput (Mbps) for Setting 1.

The time-averaged throughput results for Setting 1 are shown in Fig. 1 for two clients, while the simulation results for Setting 2 are shown in Fig. 2, where our OFTRL-LogBar-Bandit algorithm is denoted as OFTRL.



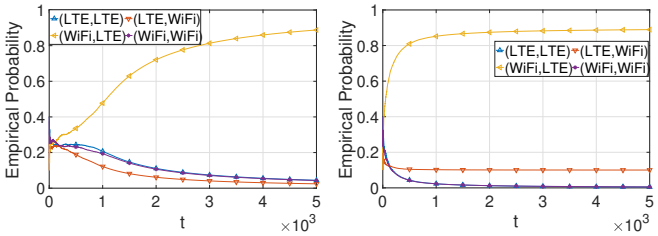
(a) Average Throughput (b) Box Plot

Figure 2: The time-averaged throughput (Mbps) for Setting 2.

Fig. 2a displays the mean time-averaged throughput across 20 clients, and Fig. 2b presents the variability in time-averaged throughput among individual clients using boxplots.

The plot for regret is omitted because time-averaged throughput effectively conveys similar information. A smaller gap from the optimal actions in hindsight (denoted as Opt) corresponds to lower regret incurred by an algorithm.

OFTRL-LogBar-Bandit demonstrates a faster convergence rate compared to other algorithms in both settings, validating the consistency of our analytical results.



(a) Lights (b) OFTRL-LogBar-Bandit

Figure 3: The empirical distribution of joint actions over time in Setting 1.

B. Convergence to Correlated Equilibrium

Fig. 3 demonstrates the convergence of empirical distributions of joint actions for Lights and OFTRL-LogBar-Bandit in Setting 1, where (WiFi, LTE) is the optimal solution when Client 1 attaches to WiFi and Client 2 attaches to LTE. It is evident that both algorithms converge towards a CE, Notably, OFTRL-LogBar-Bandit exhibits a faster convergence rate to the CE, benefiting from a swap regret that is less dependent on the number of rounds T.

VII. Conclusion

In this paper, we demonstrate that the OFTRL-LogBar-Bandit algorithm achieves the time-averaged swap regret of $\tilde{O}(T^{3/4})$, i.e., a decay rate of $\tilde{O}(T^{-3/4})$ for the time-averaged swap-regret, in unknown general-sum games.

Future work will explore whether swap regret with a time dependence comparable to that of full-information feedback can be achieved.

APPENDIX

A. Useful Facts

Definition A.1 (Directed Tree)**.** A directed graph $\mathcal{T}_a = (V, E)$ is a directed tree rooted at node a if it satisfies the following conditions:

1) it contains no (directed) cycles,
2) every node in $V \setminus a$ has exactly one outgoing edge,
3) the root node a does not have any outgoing edges.

Then, let $\mathbb{T}_a$ denote the set of all $\mathcal{T}_a$, and we are ready to state the Markov chain tree theorem as follows.

Theorem A.2 (Markov Chain Tree Theorem [40])**.** *Let $p \in \Delta^m$ be a stationary distribution for an ergodic (i.e., aperiodic and irreducible) Markov chain with the transition matrix described by Q, which can be calculated as follows:*

$$p(a) = \frac{\Sigma_a}{\Sigma}, \tag{8}$$

where $\Sigma_a := \sum_{\mathcal{T}\in\mathbb{T}_a} \prod_{(u,v)\in E(\mathcal{T})} Q(u,v)$, and $\sum := \sum_a \Sigma_a$.

B. Proof of Lemma V.1

Proof. The proof follows two steps as follows.

1) We claim that for any $t \in [T]$:
$$\sum_{a\in\mathcal{A}_n} \left\|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\right\|^2_{*,q_a^t} \le 2\sum_{a\in\mathcal{A}_n} \left|u_n^t(a) - m_n^t(a)\right| \mathbf{1}[a_n^t = a] + 2\left\|p_n^t - p_n^{t-1}\right\|_1. \tag{9}$$
2) Next, we complete the proof by demonstrating that when summed over $t \in [T]$, we have
$$\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_n} |u_n^t(a) - m_n^t(a)|\,\mathbf{1}[a_n^t = a] = \sum_{t=1} \|u_n^t - u_n^{t-1}\|_1.$$

Now, we prove our claim in Step 1 as follows. Recall that $\|x\|_{*,q_a^t} := \sqrt{\sum_{a'\in\mathcal{A}_n}(x(a')q_a^t(a'))^2}$. Then, we have

$$\begin{aligned}
&\left\|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\right\|^2_{*,q_a^t}\\
&= \sum_{a'\in\mathcal{A}_n} (q_a^t(a'))^2 \left(p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t + p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t\right)^2\\
&\le 2\sum_{a'\in\mathcal{A}_n} (q_a^t(a'))^2 \left(\left(p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t\right)^2 + \left(p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t\right)^2\right)\\
&= 2\left\|p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t\right\|^2_{*,q_a^t} + 2\left\|p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t\right\|^2_{*,q_a^t},
\end{aligned} \tag{10}$$

where the inequality is due to the Cauchy-Schwarz inequality.

Summing over $a \in \mathcal{A}_n$, the first term on the RHS of (10) can be further bounded as follows.

$$\begin{aligned}
&2\sum_{a\in\mathcal{A}_n} \left\|p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t\right\|^2_{*,q_a^t}\\
&= 2\sum_{a\in\mathcal{A}_n}\sum_{a'\in\mathcal{A}_n} (q_a^t(a')^2 \left(\frac{p_n^t(a)(u_n^t(a') - m_n^t(a'))}{p_n^t(a')}\right)^2 \mathbf{1}[a_n^t = a']\\
&\le 2\sum_{a\in\mathcal{A}_n}\sum_{a'\in\mathcal{A}_n} \frac{p_n^t(a)q_a^t(a')}{p_n^t(a')}|u_n^t(a') - m_n^t(a')|\mathbf{1}[a_n^t = a']\\
&\le 2\sum_{a'\in\mathcal{A}_n} |u_n^t(a') - m_n^t(a')|\mathbf{1}[a_n^t = a'],
\end{aligned}$$

where the first inequality is due to $\frac{p_n^t(a)q_a^t(a')}{p_n^t(a')} \le 1$ and $|u_n^t(a') - m_n^t(a')| \le 1$. The last inequality is due to the definition of p_n^t such that $p_n^t(a') = \sum_{a\in\mathcal{A}_n} p_n^t(a)q_a^t(a')$.

Regarding the second term on the RHS of (10), we have that

$$\begin{aligned}
&2\sum_{a\in\mathcal{A}_n} \left\|p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t\right\|^2_{*,q_a^t}\\
&= 2\sum_{a\in\mathcal{A}_n}\sum_{a'\in\mathcal{A}_n} (q_a^t(a'))^2 \left(p_n^t(a)m_n^t(a') - p_n^{t-1}(a)m_n^t(a')\right)^2\\
&\le 2\sum_{a\in\mathcal{A}_n}\sum_{a'\in\mathcal{A}_n} q_a^t(a')|p_n^t(a) - p_n^{t-1}(a)|\\
&= 2\|p_n^t - p_n^{t-1}\|_1,
\end{aligned}$$

where the inequality is due to that $m_n^t(a') \le 1$ and $q_a^t(a') \le 1$ for all $a' \in \mathcal{A}_n$, and the last equality is due to that $\sum_{a'\in\mathcal{A}_n} q_a^t(a') = 1$.

Then, Step 2 is obtained by summing over $t \in [T]$ and $a' \in \mathcal{A}_n$ as follows.

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_n} \left|u_n^t(a) - m_n^t(a)\right| \mathbf{1}[a_n^t = a] = \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_n} \left|u_n^t(a) - u_n^{\tau_t(a)}(a)\right| \mathbf{1}[a_n^t = a]\\
&\le \sum_{a\in\mathcal{A}_n}\sum_{t=1}^{T} \mathbf{1}[a_n^t = a] \sum_{s=\tau_t(a)+1}^{t} |u_n^s(a) - u_n^{s-1}(a)| \le \sum_{a\in\mathcal{A}_n}\sum_{t=1}^{T} |u_n^t(a) - u_n^{t-1}(a)|.
\end{aligned}$$

□

C. Proof of Lemma A.3

Lemma A.3. *When $\eta \le \frac{1}{16}$, we have that*

$$\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_n} \left\|q_a^t - q_a^{t-1}\right\|^2_{q_a^{t-1}} \ge \frac{1}{64A_n}\sum_{t=1}^{T} \left\|p_n^t - p_n^{t-1}\right\|_1^2.$$

Before we prove Lemma A.3, we need the following lemma.

Lemma A.4 (Multiplicative Stability for Bandits)**.** *For $\eta \le \frac{1}{16}$, $\|\hat{u}_n^t\|_\infty \le 1$ and $\|p_n^{t-1}(a)m_n^t\|_\infty \le 1$ for all $t \in [T]$ and $a \in \mathcal{A}_n$, we have that*

$$\sum_{a\in\mathcal{A}_n} \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}} \le \frac{1}{2}.$$

Proof of Lemma A.4. Using (10), we have that

$$\begin{aligned}
&\left\|p_n^{t-1}(a)\hat{u}_{n,a}^{t-1} - p_n^{t-2}(a)m_n^{t-1}\right\|_{*,q_a^{t-1}}\\
&\le 2\left\|p_n^{t-1}(a)\hat{u}_{n,a}^{t-1} - p_n^{t-1}(a)m_n^{t-1}\right\|_{*,q_a^{t-1}}\\
&+ 2\left\|p_n^{t-1}(a)m_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\right\|_{*,q_a^{t-1}}.
\end{aligned} \tag{11}$$

Then, the first term on the RHS of the above inequality can be bounded as follows

$$\begin{aligned}
&\|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-1}(a)m_n^{t-1}\|^2_{*,q_a^{t-1}}\\
&\le \sqrt{\sum_{a'\in\mathcal{A}_n} (q_a^{t-1}(a'))^2 \left(\frac{p_n^{t-1}(a)\mathbf{1}[a_n^{t-1} = a'](u_n^t(a') - m_n^t(a'))}{p_n^{t-1}(a')}\right)^2}\\
&\le \sqrt{\sum_{a'\in\mathcal{A}_n} \left(\frac{p_n^{t-1}(a)q_a^{t-1}(a')}{p_n^{t-1}(a')}\right)^2 \mathbf{1}[a_n^{t-1} = a']} = \frac{p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})},
\end{aligned} \tag{12}$$

where the second inequality is due to that $|u_n^t(a) - m_n^t(a)| \leq 1$. The second term can be bounded as follows

$$\begin{aligned}
&\|p_n^{t-1}(a)m_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\|_{*,q_a^{t-1}} \\
&= \sqrt{\sum_{a' \in \mathcal{A}_n} (q_a^{t-1}(a'))^2 \left(p_n^{t-1}(a)m_n^{t-1}(a') - p_n^{t-2}(a)m_n^{t-1}(a')\right)^2} \\
&\leq \sqrt{\sum_{a' \in \mathcal{A}_n} q_a^{t-1}(a') \left(p_n^{t-1}(a) - p_n^{t-2}(a)\right)^2} \\
&= p_n^{t-1}(a) - p_n^{t-2}(a).
\end{aligned} \tag{13}$$

Then, substituting (12) and (13) into (11) gives

$$\begin{aligned}
&\left\|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\right\|_{*,q_a^{t-1}} \\
&\leq 2\frac{p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})} + 2(p_n^{t-1}(a) - p_n^{t-2}(a)).
\end{aligned}$$

With the above inequality and the definition of p_n^t and q_a^t, invoking Corollary B.4 of [17] with $m_n^t(a) = u_n^{\tau_t(a)}(a)$ gives

$$\begin{aligned}
&\sum_{a \in \mathcal{A}_n} \left\|q_a^t - q_a^{t-1}\right\|_{q_a^{t-1}} \leq \sum_{a \in \mathcal{A}_n} \Big(4\eta \left\|p_n^{t-1}(a)m_n^t\right\|_{*,q_a^{t-1}} \\
&+ 2\eta \left\|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\right\|_{*,q_a^{t-1}}\Big) \\
&\leq \sum_{a \in \mathcal{A}_n} 4\eta p_n^{t-1}(a) + 4\eta \sum_{a \in \mathcal{A}_n} \frac{p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})} \\
&+ 4\eta \sum_{a \in \mathcal{A}_n} (p_n^{t-1}(a) - p_n^{t-2}(a)) \\
&= 4\eta + 4\eta + 0 = 8\eta,
\end{aligned}$$

where the last equality uses the fact that $\sum_{a \in \mathcal{A}_n} p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1}) = p_n^{t-1}(a_n^{t-1})$. The lemma follows the assumption that $\eta \leq \frac{1}{16}$. □

Then, we are ready to give the proof of Lemma A.3 as follows.

Proof of Lemma A.3. Let $\mu_a^t := \max_{a' \in \mathcal{A}_n} \left|1 - \frac{q_a^t(a')}{q_a^{t-1}(a')}\right|$. Then, it holds that

$$(\mu_a^t)^2 \leq \sum_{a' \in \mathcal{A}_n} \left(1 - \frac{q_a^t(a')}{q_a^{t-1}(a')}\right)^2 = \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2. \tag{14}$$

According to Lemma A.4, we have that $\sum_{a \in \mathcal{A}_n} \mu_a^t \leq 0.5$. By the definition of μ_a^t, we have that for any $a, a' \in \mathcal{A}_n$:

$$(1 - \mu_a^t)q_a^{t-1}(a') \leq q_a^t(a') \leq (1 + \mu_a^t)q_a^{t-1}(a').$$

By the definition of $\mathcal{T}_a$ in Definition A.1, we have that $\prod_{(u,v) \in E(\mathcal{T}_a)} Q_n^t(u,v) = \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^t(v)$, and the above inequality implies that

$$\begin{aligned}
&\prod_{u \in \mathcal{A}_n \setminus a} (1 - \mu_u^t) \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^{t-1}(v) \leq \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^t(v) \\
&\leq \prod_{u \in \mathcal{A}_n \setminus a} (1 + \mu_u^t) \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^{t-1}(v).
\end{aligned} \tag{15}$$

Since $\Sigma_a^t = \sum_{\mathcal{T} \in \mathbb{T}_a} \prod_{(u,v) \in E(\mathcal{T})} Q_n^t(u,v)$, summing up (15) over all $\mathcal{T} \in \mathbb{T}_a$ gives that

$$\Sigma_a^t \leq \Sigma_a^{t-1} \prod_{a' \in \mathcal{A}_n} (1 + \mu_a^t) \leq \Sigma_a^{t-1} \exp\left\{\sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\},$$

and that

$$\Sigma_a^t \geq \Sigma_a^{t-1} \prod_{a' \in \mathcal{A}_n} (1 - \mu_a^t) \geq \Sigma_a^{t-1} \exp\left\{-2 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\},$$

where the last inequality is due to $1 - x \geq e^{-2x}$ for any $x \in [0, \frac{1}{2}]$ and $\sum_{a \in \mathcal{A}_n} \mu_a^t \leq \frac{1}{2}$ due to Lemma A.4.

Since $\Sigma^t = \sum_{a \in \mathcal{A}_n} \Sigma_a^t$, we have the lower and upper bounds for Σ^t as follows:

$$\Sigma^{t-1} \exp\left\{-2 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} \leq \Sigma^t \leq \Sigma^{t-1} \exp\left\{\sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\}.$$

Then, we have that

$$\begin{aligned}
p_n^t(a) - p_n^{t-1}(a) &= \frac{\Sigma_a^t}{\Sigma^t} - \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \\
&\leq \frac{\exp\left\{\sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} \Sigma_a^{t-1}}{\exp\left\{-2 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} \Sigma^{t-1}} - \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \\
&= \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \left(\exp\left\{3 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} - 1\right) \\
&\leq \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \left(8 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right),
\end{aligned}$$

where the last inequality is due to $e^x - 1 \leq \frac{8}{3}x$ for all $x \in [0, \frac{3}{2}]$. Similarly, we have that

$$\begin{aligned}
p_n^{t-1}(a) - p_n^t(a) &= \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} - \frac{\Sigma_a^t}{\Sigma^t} \\
&\leq \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} - \frac{\exp\left\{-2 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} \Sigma_a^{t-1}}{\exp\left\{\sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\} \Sigma^{t-1}} \\
&= \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \left(1 - \exp\left\{-3 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right\}\right) \\
&\leq \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}} \left(3 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right),
\end{aligned}$$

where the last inequality is due to $1 - x \leq e^{-x}$. Thus, by the fact that $p_n^{t-1} = \frac{\Sigma_a^{t-1}}{\Sigma^{t-1}}$, we obtain that

$$|p_n^t(a) - p_n^{t-1}(a)| \leq p_n^{t-1}(a) \left(8 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t\right).$$

Summing over $a \in \mathcal{A}_n$ gives $\|p_n^t - p_n^{t-1}\|_1 \leq 8 \sum_{a' \in \mathcal{A}_n} \mu_{a'}^t$.

Thus, by Cauchy-Schwarz inequality and (14), we have that

$$\begin{aligned}\|p_n^t - p_n^{t-1}\|_1^2 &\le 64\left(\sum_{a\in\mathcal{A}_n}\mu_a^t\right)^2 \le 64A_n\sum_{a\in\mathcal{A}_n}(\mu_a^t)^2 \\ &\le 64A_n\sum_{a\in\mathcal{A}_n}\|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2.\end{aligned} \tag{16}$$

□

D. Proof of Theorem V.2

Proof. **Step 1:** We express the swap regret bound in terms of the estimated reward as follows. Notice that

$$\begin{aligned}&\mathbf{E}\left[\sum_{a\in\mathcal{A}_n}\langle q_a^t, p_n^t(a)\hat{u}_n^t\rangle \mid \mathcal{F}_{t-1}\right] = \sum_{a\in\mathcal{A}_n}\sum_{a'\in\mathcal{A}_n} p_n^t(a)q_a^t(a')m_n^t(a') \\ &- \mathbf{E}\left[\frac{p_n^t(a)q_a^t(a')\mathbf{1}[a_n^t = a'](m_n^t(a') - u_n^t(a'))}{p_n^t(a')} \mid \mathcal{F}_{t-1}\right] \\ &= \sum_{a'\in\mathcal{A}_n} p_n^t(a')m_n^t(a') - \sum_{a'\in\mathcal{A}_n} p_n^t(a')(m_n^t(a') - u_n^t(a')) \\ &= \sum_{a\in\mathcal{A}_t} p_n^t(a)u_n^t(a),\end{aligned}$$

and that

$$\begin{aligned}&\mathbf{E}\left[\sum_{a\in\mathcal{A}_n} p_n^t(a)\hat{u}_n^t(F(a)) \mid \mathcal{F}_{t-1}\right] = \sum_{a\in\mathcal{A}_n} p_n^t(a)m_n^t(F(a)) \\ &- \sum_{a\in\mathcal{A}_n} p_n^t(a)(m_n^t(F(a)) - u_n^t(F(a))) = \sum_{a\in\mathcal{A}_n} p_n^t(a)u_n^t(F(a)).\end{aligned}$$

Then, the regret defined in (1) can be converted as follows:

$$\begin{aligned}R_n^{\mathrm{swa}}(T) &= \max_{F\in\mathcal{F}_n}\mathbf{E}\left[\sum_{t=1}^T\sum_{a\in\mathcal{A}_n}\mathbf{1}[a_n^t = a]\left(u_n^t(F(a)) - u_n^t(a)\right)\right] \\ &= \max_{F\in\mathcal{F}_n}\mathbf{E}\left[\sum_{t=1}^T\mathbf{E}\left[\sum_{a\in\mathcal{A}_n} p_n^t(a)\left(u_n^t(F(a)) - u_n^t(a)\right]\right) \mid \mathcal{F}_{t-1}\right]\Bigg] \\ &= \max_{F\in\mathcal{F}_n}\sum_{t=1}^T\sum_{a\in\mathcal{A}_n}\mathbf{E}\left[p_n^t(a)\hat{u}_n^t(F(a)) - \langle q_a^t, p_n^t(a)\hat{u}_n^t\rangle\right].\end{aligned} \tag{17}$$

Step 2: Notice that

$$p_n^t(a)\hat{u}_n^t(F(a)) - \langle q_a^t, p_n^t(a)\hat{u}_n^t\rangle \le \underbrace{\max_{q\in\Delta(\mathcal{A}_n)}\langle q - q_a^t, p_n^t(a)\hat{u}_n^t\rangle}_{=:R_a(T)}.$$

Thus, we reduce the problem of bounding the swap regret to bounding external regret $R_a(T)$ for each subroutine $a \in \mathcal{A}_n$. This enables us to refine the existing full-information feedback analytical results to be applicable to our bandit feedback setting.

Now we want to invoke Corollary B.3. in [17] to bound $R_a(T)$, which requires $\eta\|p_n^t(a)\hat{u}_n^t - p_n^{t-1}m_n^T\|_{*,q_a^t} \le \frac{1}{8}$ and $\eta\|p_n^{t-1}m_n^t\|_{*,g_a^{t-1}} \le \frac{1}{2}$, where $g_a^t \in \Delta(\mathcal{A}_n)$ is an auxiliary probability distribution defined as the solution when the predictor is exactly the estimated reward for the t-th round for subroutine $a \in \mathcal{A}_n$.

In the following, we show that the requirement for Corollary B.3. in [17] is satisfied when $\eta \le \frac{1}{32}$. According to (11) to (13), we have that

$$\begin{aligned}&\eta\left\|p_n^t\hat{u}_n^t - p_n^{t-1}m_n^t\right\|_{*,q_a^t} \\ &\le 2\eta\left(\left\|p_n^t\hat{u}_n^t - p_n^t m_n^t\right\|_{*,q_a^t} + \left\|p_n^t m_n^t - p_n^{t-1}m_n^t\right\|_{*,q_a^t}\right) \\ &\le 2\eta\left(\frac{p_n^t(a)q_a^t(a_n^t)}{p_n^t(a_n^t)} + p_n^t(a) - p_n^{t-1}(a)\right) \le 4\eta \le \frac{1}{8},\end{aligned} \tag{18}$$

where the third inequality is due to $\frac{p_n^t(a)q_a^t(a_n^t)}{p_n^t(a_n^t)} \le 1$ and $p_n^t(a) - p_n^{t-1}(a) \le 1$, and the last inequality is due to $\eta \le \frac{1}{32}$.

On the other hand, for any $q \in \Delta(\mathcal{A}_n)$, we have that

$$\eta\left\|p_n^{t-1}m_n^t\right\|_{*,q} \le \eta\left\|p_n^{t-1}m_n^t\right\|_\infty \le \eta \le \frac{1}{2}.$$

Then, we can invoke Corollary B.3. in [17] to bound $R_a(T)$ as follows:

$$\begin{aligned}R_a(T) \le \frac{2A_n\ln T}{\eta} &+ 2\eta\sum_{t=1}^T\left\|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\right\|_{*,q_a^t}^2 \\ &- \frac{1}{16\eta}\sum_{t=1}^T\left\|q_a^t - q_a^{t-1}\right\|_{q_a^{t-1}}^2.\end{aligned}$$

Step 3: Summing over $a \in \mathcal{A}_n$ and invoking Lemmas V.1 and A.3, we have the swap regret bounded as follows:

$$\begin{aligned}\sum_{a\in\mathcal{A}_n} R_a(T) \le \frac{2(A_n)^2\ln T}{\eta} &+ 2\eta\sum_{t=1}^T\|u_n^t - u_n^{t-1}\|_1 + 2\eta\sum_{t=1}^T\|p_n^t - p_n^{t-1}\|_1. \\ &- \frac{1}{1024A_n\eta}\sum_{t=1}^T\|p_n^t - p_n^{t-1}\|_1^2.\end{aligned}$$

Next, for any $t \in [T]$ and $a \in \mathcal{A}_n$, since $u_n^t(a) = \mathbf{E}_{A_{-n}\sim p_{-n}^t}[u_n(a; A_{-n}^t)] = \sum_{A_{-n}} p_{-n}^t(A_{-n})u_n(a; A_{-n}^t)$, we have that

$$\begin{aligned}&|u_n^t(a) - u_n^{t-1}(a)| \\ &= \left|\sum_{A_{-n}} p_{-n}^t(A_{-n})u_n(a; A_{-n}) - \sum_{A_{-n}} p_{-n}^{t-1}(A_{-n})u_n(a; A_{-n})\right| \\ &\le \sum_{A_{-n}}|u_n(a; A_{-n})|\left|p_{-n}^t(A_{-n}) - p_{-n}^{t-1}(A_{-n})\right| \\ &\le \sum_{A_{-n}}\left|\prod_{m\ne n} p_m^t[a_m] - \prod_{m\ne n} p_m^{t-1}[a_m]\right| \le \sum_{m\ne n}\left\|p_m^t - p_m^{t-1}\right\|_1,\end{aligned}$$

where the last inequality is due to the total distance between two product distributions being bounded by the sum of the total variations of each marginal distribution [35].

Then, we have the swap regret bounded as follows:

$$\begin{aligned}R_n^{\mathrm{swa}}(T) \le \frac{2(A_n)^2\ln T}{\eta} &+ 2\eta\sum_{t=1}^T\sum_{m\in\mathcal{A}_n}\left\|p_m^t - p_m^{t-1}\right\|_1 \\ &- \frac{1}{1024A_n\eta}\sum_{t=1}^T\left\|p_n^t - p_n^{t-1}\right\|_1^2.\end{aligned}$$

□

References

[1] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[2] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, July 2020.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Non-Stochastic Multiarmed Bandit Problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[4] A. Blum and Y. Mansour, "From External to Internal Regret," *Journal of Machine Learning Research (JMLR)*, vol. 8, no. 6, 2007.

[5] S. Ito, "A Tight Lower Bound and Efficient Reduction for Swap Regret," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18 550–18 559.

[6] Z. Huang and J. Pan, "A Near-Optimal High-Probability Swap-Regret Upper Bound for Multi-Agent Bandits in Unknown General-Sum Games," in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, vol. 216. PMLR, 31 Jul–04 Aug 2023, pp. 911–921.

[7] R. Karp, E. Koutsoupias, C. Papadimitriou, and S. Shenker, "Optimization Problems in Congestion Control," in *Proc. Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2000, pp. 66–74.

[8] C. Papadimitriou, "Algorithms, Games, and the Internet," in *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, 2001, pp. 749–753.

[9] Z. Huang and J. Pan, "End-to-End Congestion Control as Learning for Unknown Games with Bandit Feedback," in *Proc. IEEE Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2023.

[10] ——, "Distributed Learning of Unknown Games for HetNet Selection," *IEEE Transactions on Network Science and Engineering (TNSE)*, 2024.

[11] D. D. Nguyen, H. X. Nguyen, and L. B. White, "Reinforcement Learning with Network-Assisted Feedback for Heterogeneous RAT Selection," *IEEE Transactions on Wireless Communications (TWC)*, vol. 16, no. 9, pp. 6062–6076, 2017.

[12] X. Li, Q. Huang, and D. Wu, "A Repeated Stochastic Game Approach for Strategic Network Selection in Heterogeneous Networks," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2018, pp. 88–93.

[13] H. H. Nax, M. N. Burton-Chellew, S. A. West, and H. P. Young, "Learning in a Black Box," *Journal of Economic Behavior & Organization*, vol. 127, pp. 1–15, 2016.

[14] R. J. Aumann, "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.

[15] C.-Y. Wei and H. Luo, "More Adaptive Algorithms for Adversarial Bandits," in *Proc. Conference on Learning Theory (COLT)*. PMLR, 2018, pp. 1263–1291.

[16] C. Jin, Q. Liu, Y. Wang, and T. Yu, "V-Learning—A Simple, Efficient, Decentralized Algorithm for Multiagent RL," in *Proc. ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

[17] I. Anagnostides, G. Farina, C. Kroer, C.-W. Lee, H. Luo, and T. Sandholm, "Uncoupled Learning Dynamics with $O(\log T)$ Swap Regret in Multiplayer Games," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 3292–3304.

[18] G. W. Brown, "Some Notes on Computation of Games Solutions," RAND Corp Santa Monica CA, Tech. Rep., 1949.

[19] J. Robinson, "An Iterative Method of Solving a Game," *Annals of Mathematics*, pp. 296–301, 1951.

[20] B. Peng and A. Rubinstein, "Fast Swap Regret Minimization and Applications to Approximate Correlated Equilibria," in *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, 2024, pp. 1223–1234.

[21] Y. Dagan, C. Daskalakis, M. Fishelson, and N. Golowich, "From External to Swap Regret 2.0: An Efficient Reduction for Large Action Spaces," in *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, 2024, pp. 1216–1222.

[22] W. Krichene, B. Drighès, and A. M. Bayen, "Online Learning of Nash Equilibria in Congestion Games," *SIAM Journal on Control and Optimization*, vol. 53, no. 2, pp. 1056–1081, 2015.

[23] G. Palaiopanos, I. Panageas, and G. Piliouras, "Multiplicative Weights Update with Constant Step-Size in Congestion Games: Convergence, Limit Cycles and Chaos," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5872–5882.

[24] X. Chen and B. Peng, "Hedging in Games: Faster Convergence of External and Swap Regrets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] C. Daskalakis, M. Fishelson, and N. Golowich, "Near-Optimal No-Regret Learning in General Games," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[26] I. Anagnostides, C. Daskalakis, G. Farina, M. Fishelson, N. Golowich, and T. Sandholm, "Near-Optimal No-Regret Learning for Correlated Equilibria in Multi-Player General-Sum Games," in *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, 2022, pp. 736–749.

[27] G. Farina, I. Anagnostides, H. Luo, C.-W. Lee, C. Kroer, and T. Sandholm, "Near-Optimal No-Regret Learning Dynamics for General Convex Games," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[28] G. Stoltz, "Incomplete Information and Internal Regret in Prediction of Individual Sequences," Ph.D. dissertation, Université Paris Sud-Paris XI, 2005.

[29] Y. Cai, A. Oikonomou, and W. Zheng, "Finite-Time Last-Iterate Convergence for Learning in Multi-Player Games," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 33 904–33 919.

[30] Y. Cai, G. Farina, J. Grand-Clément, C. Kroer, C.-W. Lee, H. Luo, and W. Zheng, "Fast Last-Iterate Convergence of Learning in Games Requires Forgetful Algorithms," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[31] R. P. Leme, G. Piliouras, and J. Schneider, "Convergence of No-Swap-Regret Dynamics in Self-Play," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[32] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, "Fast Convergence of Regularized Learning in Games," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[33] T. Kocák, G. Neu, M. Valko, and R. Munos, "Efficient Learning by Implicit Exploration in Bandit Problems with Side Observations," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 613–621.

[34] G. Neu, "Explore No More: Improved High-Probability Regret Bounds for Non-Stochastic Bandits," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[35] W. Hoeffding and J. Wolfowitz, "Distinguishability of Sets of Distributions," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 700–718, 1958.

[36] Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos, "Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium," in *Proc. Conference on Learning Theory (COLT)*. PMLR, 2021, pp. 2388–2422.

[37] B. N. Feinberg and S. S. Chiu, "A Method to Calculate Steady-State Distributions of Large Markov Chains by Aggregating States," *Operations Research*, vol. 35, no. 2, pp. 282–290, 1987.

[38] M. B. Cohen, J. Kelner, J. Peebles, R. Peng, A. B. Rao, A. Sidford, and A. Vladu, "Almost-Linear-Time Algorithms for Markov Chains and New Spectral Primitives for Directed Graphs," in *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, 2017, pp. 410–419.

[39] S. Hart and A. Mas-Colell, "A Reinforcement Procedure Leading to Correlated Equilibrium," in *Economics Essays: A Festschrift for Werner Hildenbrand*. Springer, 2001, pp. 181–200.

[40] V. Anantharam and P. Tsoucas, "A Proof of the Markov Chain Tree Theorem," *Statistics & Probability Letters*, vol. 8, no. 2, pp. 189–192, 1989.