

# Faster Convergence for Unknown-Game Bandits

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  $u_n : \mathcal{A} \rightarrow [0, 1]$ , 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  $\mathcal{F}_n := \{F : \mathcal{A}_n \rightarrow \mathcal{A}_n\}$ , which results in the following definition of *swap regret*:

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

We can reduce swap regret to external regret by restricting  $\mathcal{F}_n$  to the subset  $\{F_a, \forall a \in \mathcal{A}_n : \mathcal{A}_n \rightarrow a\}$ . 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  $\epsilon$ -correlated equilibria (CE) [1], [4], where  $A^t := (a_1^t, \dots, a_N^t)$  is the joint actions played by all agents in round  $t$ . The notion of  $\epsilon$ -CE is defined as follows.

**Definition III.1** ( $\epsilon$ -Correlated Equilibrium [14]). A joint probability distribution  $\mathbf{P}$  over  $\mathcal{A}$  is an  $\epsilon$ -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

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

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  $\epsilon$ -CE. The primary objective of this work is to achieve a faster convergence rate for swap regret, i.e., minimizing the dependence of  $R_n^{\text{swa}}(T)$  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

---

```

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

$$(p_n^t)^\top = (p_n^{t-1})^\top Q_n^t, \quad (3)$$

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:

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

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  $\tau_t(a) := \arg \max_{s < t} \{1[a_n^s = a]\}$  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:

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

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 local norm of vector  $x \in \mathbb{R}^{A_n}$  with regard to the log-barrier function with value  $q_a^t$ , 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  $\sigma$ -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 plays 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*

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

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

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

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

- 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:

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

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

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  $\eta \leq \frac{1}{16}$ , the swap regret for each agent  $n \in [N]$  is upper bounded as follows.*

$$\begin{aligned} R_n^{\text{swa}}(T) & \leq 2 + \mathbf{E} \left[ \frac{2(A_n)^2 \ln T}{\eta} + 4\eta \sum_{t=1}^T \sum_{m \in [N]} \|p_m^t - p_m^{t-1}\|_1 \right. \\ & \quad \left. - \frac{1}{1024A_n\eta} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2 \right]. \end{aligned}$$

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

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

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

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:

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

We want to leverage Corollary B.3 of [17] (for completeness, we reproduce their results in Lemma A.10) to bound the RHS of the above equation. However, Lemma A.10 requires  $q \in \text{relint}(\Delta(\mathcal{A}_n))$ , while  $q \in \Delta(\mathcal{A}_n)$  for  $R_a^T(q)$ . Thus, we need to find a surrogate point  $\tilde{q}_a \in \text{relint}(\Delta(\mathcal{A}_n))$  such that  $R_a^T(q) = R_a^T(\tilde{q}_a) +$  some penalty terms.



Let  $q_a^* = \arg \max_{q \in \Delta(\mathcal{A}_n)} \mathbf{E}[R_a^T(q)]$ , and  $q_a^c(a') = \frac{1}{A_n}, \forall a' \in \mathcal{A}_n$ . Now, we define  $\tilde{q}_a = (1 - \frac{1}{T}) q_a^* + \frac{1}{T} q_a^c$ , and thus  $\tilde{q}_a \in \text{relint}(\Delta(\mathcal{A}_n))$ . With  $\tilde{q}_a$ , we have  $R_a^T(q_a^*)$  bounded by

$$\begin{aligned} \mathbf{E}[R_a^T(q_a^*)] &= \mathbf{E}\left[\sum_{t=1}^T \frac{1}{T} \langle q_a^* - q_a^c, p_n^t(a) \hat{u}_n^t \rangle\right] + \mathbf{E}[R_a^T(\tilde{q}_a)] \\ &\leq \mathbf{E}\left[\sum_{t=1}^T \frac{1}{T} \|q_a^* - q_a^c\|_1 \|p_n^t(a) \hat{u}_n^t\|_\infty\right] + \mathbf{E}[R_a^T(\tilde{q}_a)] \\ &\leq \mathbf{E}\left[\sum_{t=1}^T \frac{2}{T} p_n^t(a)\right] + \mathbf{E}[R_a^T(\tilde{q}_a)], \end{aligned}$$

where the first inequality is due to Holder's inequality, and the last inequality is due to  $\|q_a^* - q_a^c\|_1 \leq \|q_a^*\|_1 + \|q_a^c\|_1 \leq 2$  and  $u_n^t \leq 1$ , and the RHS of the above equation can be bounded by invoking Lemma A.10 as follows:

$$\begin{aligned} \mathbf{E}[R_a^T(\tilde{q}_a)] &\leq \mathbf{E}\left[\frac{2A_n \ln T}{\eta} + 2\eta \sum_{t=1}^T \|p_n^t(a) \hat{u}_n^t - p_n^{t-1}(a) m_n^t\|_{*, q_a^t}^2\right. \\ &\quad \left. - \frac{1}{16\eta} \sum_{t=1}^T \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2\right]. \end{aligned} \quad (7)$$

While leveraging Lemma A.10, 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 Lemma A.10 still hold with this adapted vector for bandit feedback. Second, Lemma A.10 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.11. 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.11 (see Appendix D) to bound that  $\sum_{t=1}^T \sum_{a \in \mathcal{A}_n} \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2 \geq \frac{1}{64A_n} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2$ . Then, we obtain

$$\begin{aligned} \sum_{a \in \mathcal{A}_n} R_a^T &\leq 2 + \mathbf{E}\left[\frac{2(A_n)^2 \ln T}{\eta} + 4\eta \sum_{t=1}^T \|u_n^t - u_n^{t-1}\|_1\right. \\ &\quad \left.+ 4\eta \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1 - \frac{1}{1024A_n\eta} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2\right]. \end{aligned}$$

Theorem V.2 follows by proving  $\|u_n^t - u_n^{t-1}\|_1 \leq \sum_{m \in [N] \setminus n} \|p_m^t - p_m^{t-1}\|_1$ , 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*

$$\sum_{n \in [N]} R_n^{\text{swa}}(T) \leq 192N^{\frac{3}{2}}(A_{\max})^2 T^{\frac{1}{4}} (\ln T)^{\frac{3}{4}} + 2N.$$

*Proof.* Since  $\eta \leq \frac{1}{16}$ , invoking Theorem V.2 gives

$$\begin{aligned} \sum_{n \in [N]} R_n^{\text{swa}}(T) &\leq 2N + \frac{2N(A_{\max})^2 \ln T}{\eta} + \mathbf{E}\left[4\eta N \sum_{t=1}^T \sum_{n \in [N]} \|p_n^t - p_n^{t-1}\|_1\right. \\ &\quad \left.- \frac{1}{1024A_{\max}\eta} \sum_{n \in [N]} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2\right]. \end{aligned}$$

Let  $x_n^t := \|p_n^t - p_n^{t-1}\|_1$ , and we obtain the following upper bound on the total swap regret, which takes the form of a quadratic function

$$\begin{aligned} \sum_{n \in [N]} R_n^{\text{swa}}(T) &\leq 2N + \frac{2N(A_{\max})^2 \ln T}{\eta} + \mathbf{E}\left[4096\eta^3 A_{\max} N^3 T\right. \\ &\quad \left.- \frac{1}{1024\eta A_{\max}} \sum_{t=1}^T \sum_{n \in [N]} (x_n^t - 2048\eta^2 A_{\max} N)^2\right] \\ &\leq 2N + \frac{2N(A_{\max})^2 \ln T}{\eta} + 4096\eta^3 A_{\max} N^3 T. \end{aligned}$$

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 upper bound of sum of the swap regret for all  $N$  agents is  $\tilde{O}\left(N^{\frac{3}{2}}(A_{\max})^2 T^{\frac{1}{4}}\right)$ , where  $\tilde{O}(\cdot)$  hides the logarithmic factors. This implies that the time-averaged swap regret for all agents is  $\tilde{O}\left(N^{\frac{3}{2}}(A_{\max})^2 T^{-\frac{3}{4}}\right)$ , 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 of at least  $\tilde{O}(T^{\frac{3}{4}})$ . 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  $\tilde{O}(T^{-\frac{3}{4}})$ . Since we consider the expected reward over the opponent's mixed actions, it will be interesting to see whether the convergence rate for the full bandit feedback can be 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 OFTRL-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	
Client 1	LTE	LTE	WiFi
	WiFi	(17.5, 17.5)	(35, 24)
		(48, 35)	(16, 16)

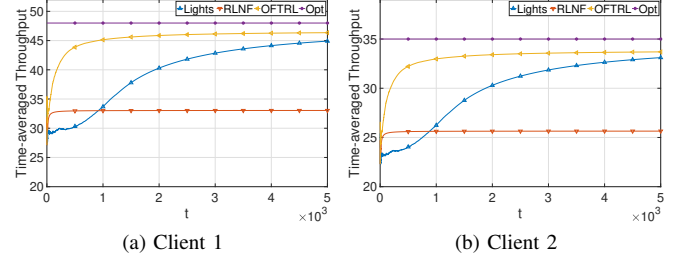


Figure 1: The time-averaged throughput (Mbps) for Setting 1.

### A. Time-averaged Throughput

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.

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.

### 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.

## REFERENCES

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

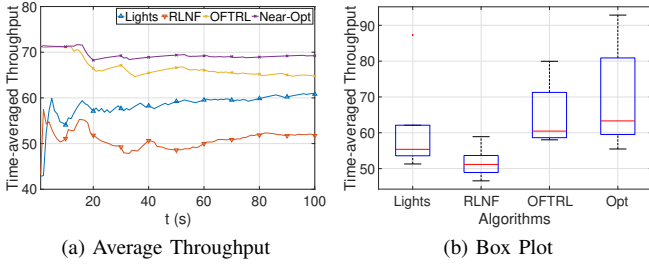


Figure 2: The time-averaged throughput (Mbps) for Setting 2.

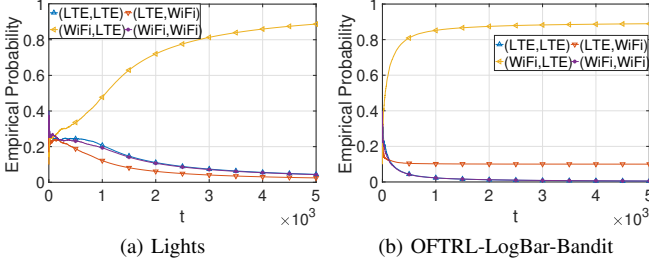


Figure 3: The empirical distribution of joint actions over time in Setting 1.

[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] A. Nemirovski, "Interior Point Polynomial Time Methods in Convex Programming," *Lecture notes*, vol. 42, no. 16, pp. 3215–3224, 2004.
- [41] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013, vol. 87.
- [42] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics, 1994.
- [43] V. Anantharam and P. Tsoucas, "A Proof of the Markov Chain Tree Theorem," *Statistics & Probability Letters*, vol. 8, no. 2, pp. 189–192, 1989.



## APPENDIX

### A. Useful Facts

a) *Self-Concordant functions.*: A concept closely tied to local norms is that of *self-concordant functions*, a special class of convex functions whose curvature changes are smoothly controlled. Self-concordant functions were introduced by Nesterov and Nemirovski in the context of interior-point methods [40]–[42]. The key idea is that the third-order directional derivative in any direction is bounded by the cube of the local norm, ensuring well-behaved curvature for optimization.

**Definition A.1** (Self-Concordant Function). A convex and three-times differentiable function  $\psi : \text{dom}(\psi) \rightarrow \mathbb{R}$  is said to be *self-concordant* on a nonempty, open, convex set  $\text{dom}(\psi) \subseteq \mathbb{R}^d$  if for all  $p \in \text{dom}(\psi)$  and all directions  $h \in \mathbb{R}^d$ , it satisfies:

$$|D^3\psi(p)[h, h, h]| \leq 2 (h^\top \nabla^2 \psi(p) h)^{3/2} = 2 \|h\|_{\nabla^2 \psi(p)}^3,$$

where  $D^3\psi(p)[h, h, h]$  denotes the third directional derivative of  $\psi$  at  $p$  in direction  $h$ , and  $\|h\|_{\nabla^2 \psi(p)} := \sqrt{h^\top \nabla^2 \psi(p) h}$  is the local norm induced by  $\psi$  at  $p$ .

Furthermore, we call  $\psi$  a  $\theta$ -self-concordant barrier, if for all  $p \in \text{dom}(\psi)$  and  $h \in \mathbb{R}^d$ ,

$$|\langle \nabla \psi(p), h \rangle|^2 \leq \theta \|h\|_{\nabla^2 \psi(p)}^2. \quad (8)$$

We provide an example of self-concordant functions as follows.

*Example A.1* (Log-Barrier). A canonical example of a self-concordant function is the log-barrier function defined on the positive orthant  $\mathbb{R}_{>0}^d$  by

$$\psi(p) = - \sum_{a \in [d]} \ln p(a).$$

This function is convex and diverges to  $+\infty$  as any  $p(a) \rightarrow 0^+$ , making it a barrier function. Its Hessian is diagonal with entries  $\nabla^2 \psi(p)_{aa} = \frac{1}{p(a)^2}$ , and its third-order directional derivative satisfies

$$|D^3\psi(p)[h, h, h]| = 2 (h^\top \nabla^2 \psi(p) h)^{3/2},$$

So it meets the self-concordance condition with equality. In addition,  $\psi(p)$  is a  $d$ -self-concordant barrier for  $p \in \mathbb{R}_{>0}^d$  according to (8), because

$$|\langle \nabla \psi(p), h \rangle|^2 = \left| \sum_{i=1}^d \left( -\frac{1}{p_i} \right) h_i \right|^2 = \left| \sum_{i=1}^d \frac{h_i}{p_i} \right|^2 \leq \left( \sum_{i=1}^d \frac{h_i^2}{p_i^2} \right) \cdot d = d \cdot \|h\|_{\nabla^2 \psi(p)}^2,$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Now, we present a lemma from (2.3.3) in Proposition 2.3.2 of that bounds the difference in function values of a self-concordant barrier based on a certain notion of distance between two interior points.

**Lemma A.2.** For any  $\theta$ -self-concordant barrier  $\psi$  and  $p, q \in \text{int}(\text{dom}(\psi))$ , we have

$$\psi(p) - \psi(q) \leq \theta \ln \left( \frac{1}{1 - \pi(p; q)} \right),$$

where

$$\pi(p; q) = \inf \{ s \geq 0 : q + s^{-1}(p - q) \in \text{dom}(\psi) \}. \quad (9)$$

**Definition A.3** (Fenchel Conjugate). For a function  $\psi : \mathbb{R}^d \rightarrow [-\infty, \infty]$ , we define the Fenchel conjugate  $\psi^* : \mathbb{R}^d \rightarrow [-\infty, \infty]$  as

$$\psi^*(\theta) = \sup_{p \in \mathbb{R}^d} \langle \theta, p \rangle - \psi(p)$$

The definition of Fenchel conjugate directly leads to Fenchel-Young's inequality:

$$\langle \theta, p \rangle \leq \psi^*(\theta) + \psi(p), \forall \theta, p \in \mathbb{R}^d. \quad (10)$$

The following example will be used in our analysis for the Fenchel conjugate of quadratic norms.

*Example A.2* (Fenchel Conjugate of Quadratic Norms). Let  $A \in \mathbb{R}^{d \times d}$  be a positive definite matrix, and we define the quadratic norm for any  $p \in \mathbb{R}^d$  as  $\|p\|_A^2 := p^\top A p$ . The Fenchel conjugate of  $\frac{1}{2} \|p\|_A^2$  is thus:

$$\langle \theta, p \rangle - p^\top A p \leq \frac{1}{2} \theta^\top A^{-1} \theta = \frac{1}{2} \|\theta\|_{A^{-1}}^2.$$

Furthermore, according to Fenchel-Young's inequality in (10), we have that

$$\langle \theta, p \rangle \leq \frac{1}{2} \|\theta\|_{A^{-1}}^2 + \frac{1}{2} \|p\|_A^2, \forall \theta, p \in \mathbb{R}^d. \quad (11)$$

Next, we introduce two lemmas from [41]. The first lemma is adapted from Theorem 4.1.6 of [41], showing that a Hessian stability property of self-concordant functions:

**Lemma A.4.** *Let  $\psi$  be a self-concordant function. Then, for any  $p, \tilde{p} \in \text{dom}(\psi)$ :*

$$\nabla^2 \psi(p) \preceq \frac{1}{(1 - \|p - \tilde{p}\|_{\nabla^2 \psi(\tilde{p})})^2} \nabla^2 \psi(\tilde{p}), \quad \|p - \tilde{p}\|_{\nabla^2 \psi(\tilde{p})} < 1. \quad (12)$$

The second lemma from Theorem 4.1.7 of [41] shows that a self-concordant function (see Definition A.1) has a property that is similar in spirit to strongly convexity:

**Lemma A.5.** *Let  $\psi$  be a self-concordant function. Then, for any  $p, \tilde{p} \in \text{dom}(\psi)$ :*

$$\psi(\tilde{p}) \geq \psi(p) + \langle \nabla \psi(p), \tilde{p} - p \rangle + \omega(\|\tilde{p} - p\|_{\nabla^2 \psi(p)}),$$

where  $w(s) := s - \log(1 + s)$ .

In addition, for any  $s \in [0, 1]$ ,  $w(s)$  satisfies that

$$w(s) \geq \frac{s^2}{4}. \quad (13)$$

Another important inequality regarding the self-concordant function is given by the following lemma from (2.20) and (2.21) of [40].

**Lemma A.6.** *If  $\psi$  is a self-concordant function with  $p^* := \arg \min \psi(p)$ , then for some  $p \in \text{dom}(\psi)$  such that  $\|\nabla \psi(p)\|_{(\nabla^2 \psi(p))^{-1}} \leq \frac{1}{2}$ , we have*

$$\begin{aligned} \|p - p^*\|_{\nabla^2 \psi(p)} &\leq 2 \|\nabla \psi(p)\|_{(\nabla^2 \psi(p))^{-1}}, \\ \|p - p^*\|_{\nabla^2 \psi(p^*)} &\leq 2 \|\nabla \psi(p)\|_{(\nabla^2 \psi(p))^{-1}}. \end{aligned}$$

**Definition A.7** (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.8** (Markov Chain Tree Theorem [43]). *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:*

$$p(a) = \frac{\Sigma_a}{\Sigma}, \quad (14)$$

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

#### B. Optimistic Follow-The-Regularized-Leader with Log-Barrier Regularizer

Consider a full-information game, where the learner can observe the reward vector  $u^t > 0$  in each round  $t$ . The main idea of OFTRL is to add vector  $m^t$  as a predictor of the current-round reward before observing it:

$$p^t \in \arg \max_{p \in \Omega} \eta \left\langle m^t + \sum_{s=1}^{t-1} u^s, p \right\rangle - \psi(p).$$

Thus, if we had perfect predictions, i.e., if  $m^t = y^t$ , we could select the optimal action for the current round. However, in practice, accurately predicting the actual loss is generally infeasible. Despite this, even when the prediction is inaccurate, the OFTRL algorithm retains its worst-case performance guarantee.

The following result is adapted from Corollary B.3 of [17], which demonstrates that OFTRL with a self-concordant regularizer satisfies the desired property. For completeness, we reproduce their proof below.

**Lemma A.9.** *Assume  $\psi$  is a nondegenerate (i.e.,  $\nabla^2 \psi(p)$  is positive definite) self-concordant function in  $\text{int}(\Omega)$  satisfying that  $\psi(p) \geq 0, \forall p \in \text{int}(\Omega)$  and  $\nabla^2 \psi(\tilde{p}) \preceq 2 \nabla^2 \psi(p)$  for any  $p, \tilde{p} \in \text{int}(\Omega)$  with  $\|p - \tilde{p}\|_{\nabla^2 \psi(\tilde{p})} \leq \frac{1}{4}$ . Let*

---

**Algorithm 2** Optimistic Follow-The-Regularized-Leader
 

---

- 1: **Input:**  $\psi, \eta_t > 0, \forall t \in [T]$
  - 2: **for**  $t = 1, \dots, T$  **do**
  - 3:   Output  $p^t \in \arg \max_{p \in \Omega} \eta \left\langle m^t + \sum_{s=1}^{t-1} u^s, p \right\rangle - \psi(p)$
  - 4:   Observe a reward vector  $u^t$
  - 5: **end for**
- 

$g^t := \arg \max_{g \in \Omega} \eta \left\langle \sum_{s=1}^t u^s, g \right\rangle - \psi(g)$  be the auxiliary sequence. If  $\eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}} \leq \frac{1}{8}$  and  $\eta \|m^t\|_{(\nabla^2 \psi(g^{t-1}))^{-1}} \leq \frac{1}{2}$  for all  $t \in [T]$ , then for any reward sequence  $u^1 \dots u^T \geq 0$  and any  $p \in \text{int}(\Omega)$ , we have

$$\sum_{t=1}^T \langle p - p^t, u^t \rangle \leq \frac{\psi(p)}{\eta} + 2\eta \sum_{t=1}^T \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}^2 - \frac{1}{16} \eta \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2 \psi(p^{t-1})}^2.$$

*Proof.* The proof is divided into the following steps.

- 1) We can break the one-step regret as follows:

$$\langle p - p^t, u^t \rangle = \underbrace{\langle p - g^t, u^t \rangle + \langle g^t - p^t, m^t \rangle}_{=:(a)} + \underbrace{\langle g^t - p^t, u^t - m^t \rangle}_{=:(b)}. \quad (15)$$

In the first step, we will use induction to bound (a) in (15) for any  $t \in [T]$  and  $p \in \text{int}(\Omega)$  as follows.

$$\begin{aligned} \sum_{t=1}^T \langle p - g^t, u^t \rangle + \langle g^t - p^t, m^t \rangle &\leq \frac{\psi(p)}{\eta} - \frac{1}{\eta} \sum_{t=1}^T w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) \\ &\quad - \frac{1}{\eta} \sum_{t=1}^T w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})}), \end{aligned} \quad (16)$$

where  $w(s) := s - \log(1 + s)$ .

- 2) Then, by Holder's inequality, we have (b) in (15) bounded as follows.

$$\langle g^t - p^t, u^t - m^t \rangle \leq \|g^t - p^t\|_{\nabla^2 \psi(p^t)} \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}.$$

In this step, we will further show that  $\|g^t - p^t\|_{\nabla^2 \psi(p^t)} \leq 2\eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}$ . Hence, we will obtain that

$$\langle g^t - p^t, u^t - m^t \rangle \leq 2\eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}^2.$$

- 3) The final step will give a lower bound to the last two terms in (16) as follows.

$$\sum_{t=1}^T w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})}) \geq \frac{1}{16} \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2 \psi(p^{t-1})}^2.$$

The proof follows by combining the three steps outlined above. We now provide the details for each step.

a) *Step 1: Prove (16).*: When  $T = 0$ , (16) holds trivially because  $\psi(p) \geq 0, \forall p \in \text{int}(\Omega)$ . Now, we assume (16) holds for  $T$ , and prove the case for  $T + 1$ . Let  $p = g^T$ , and since (16) holds for  $T$ , we have that

$$\begin{aligned} \sum_{t=1}^T \langle g^t - p^t, m^t \rangle - \langle g^t, u^t \rangle &\leq - \sum_{t=1}^T \langle g^t, u^t \rangle + \frac{\psi(g^T)}{\eta} \\ &\quad - \frac{1}{\eta} \sum_{t=1}^T (w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})})). \end{aligned}$$

Now, adding the term  $\langle g^{T+1} - p^{T+1}, m^{T+1} \rangle - \langle g^{T+1}, u^{T+1} \rangle$  on both sides gives

$$\begin{aligned} \sum_{t=1}^{T+1} \langle g^t - p^t, m^t \rangle - \langle g^t, u^t \rangle &\leq - \sum_{t=1}^T \langle g^t, u^t \rangle + \frac{\psi(g^T)}{\eta} + \langle g^{T+1} - p^{T+1}, m^{T+1} \rangle \\ &\quad - \langle g^{T+1}, u^{T+1} \rangle - \frac{1}{\eta} \sum_{t=1}^T (w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})})). \end{aligned} \quad (17)$$

Denote by  $\Psi^T(g) := \eta \left\langle \sum_{s=1}^T u^s, g \right\rangle - \psi(g)$ . By the definition of  $g^T = \arg \max_{g \in \text{int}(\Omega)} \Psi^T(g)$  and the first-order optimality, we have that  $\nabla \Psi^T(g^T) = 0$ . Invoking Lemma A.5 with  $-\Psi^T$  being the self-concordant function, we have that

$$-\Psi^T(p^{T+1}) + \Psi^T(g^T) \geq w(\|p^{T+1} - g^T\|_{\nabla^2 \Psi^T(g^T)}) = w(\|p^{T+1} - g^T\|_{\nabla^2 \psi(g^T)}).$$

This means that

$$\left\langle g^T, \sum_{t=1}^T u^t \right\rangle - \frac{\psi(g^T)}{\eta} \geq \left\langle p^{T+1}, \sum_{t=1}^T u^t \right\rangle - \frac{\psi(p^{T+1})}{\eta} + \frac{1}{\eta} w(\|p^{T+1} - g^T\|_{\nabla^2 \psi(g^T)}).$$

Substitute the above inequality to (17) gives

$$\begin{aligned} \sum_{t=1}^{T+1} \langle g^t - p^t, m^t \rangle - \langle g^t, u^t \rangle &\leq - \left\langle p^{T+1}, m^{T+1} + \sum_{t=1}^T u^t \right\rangle + \frac{\psi(p^{T+1})}{\eta} \\ &+ \langle g^{T+1}, m^{T+1} - u^{T+1} \rangle - \frac{1}{\eta} \sum_{t=1}^T (w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})})) \\ &- \frac{1}{\eta} w(\|p^{T+1} - g^T\|_{\nabla^2 \psi(g^T)}). \end{aligned} \quad (18)$$

On the other hand, denote by  $\Phi^t(p) := \eta \left\langle m^t + \sum_{s=1}^{t-1} u^s, p \right\rangle - \psi(p)$  be the objection function for OFTRL. Since  $p^{T+1} := \arg \max_{p \in \text{int}(\Omega)} \Phi^{T+1}(p)$ , we have  $\nabla \Phi^{T+1}(p^{T+1}) = 0$  by the first-order optimality condition. Invoking Lemma A.5 with  $-\Phi^{T+1}$  being the self-concordant function, we have that

$$-\Phi^{T+1}(g^{T+1}) \geq -\Phi^{T+1}(p^{T+1}) + w(\|p^{T+1} - g^{T+1}\|_{\nabla^2 \psi(p^{T+1})}),$$

which gives

$$\begin{aligned} \left\langle p^{T+1}, m^{T+1} + \sum_{t=1}^T u^t \right\rangle - \frac{\psi(p^{T+1})}{\eta} &\geq \left\langle g^{T+1}, m^{T+1} + \sum_{t=1}^T u^t \right\rangle - \frac{\psi(g^{T+1})}{\eta} \\ &+ \frac{1}{\eta} w(\|p^{T+1} - g^{T+1}\|_{\nabla^2 \psi(p^{T+1})}). \end{aligned}$$

Then, substitute the above inequality into (17), and we have

$$\begin{aligned} \sum_{t=1}^{T+1} \langle g^t - p^t, m^t \rangle - \langle g^t, u^t \rangle &\leq - \left\langle g^{T+1}, \sum_{t=1}^{T+1} u^t \right\rangle + \frac{\psi(g^{T+1})}{\eta} \\ &- \frac{1}{\eta} \sum_{t=1}^{T+1} (w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})})) \end{aligned} \quad (19)$$

Now, because  $\Psi^{T+1}(g^{T+1}) \geq \Psi^{T+1}(p)$  for any  $p \in \text{int}(\Omega)$ , adding  $\left\langle p, \sum_{t=1}^{T+1} u^t \right\rangle$  to both sides of the above inequality, we have

$$\begin{aligned} \sum_{t=1}^{T+1} \langle p - g^t, u^t \rangle + \langle g^t - p^t, m^t \rangle &\leq \frac{\psi(p)}{\eta} - \frac{1}{\eta} \sum_{t=1}^{T+1} w(\|p^t - g^t\|_{\nabla^2 \psi(p^t)}) \\ &- \frac{1}{\eta} \sum_{t=1}^{T+1} w(\|p^t - g^{t-1}\|_{\nabla^2 \psi(g^{t-1})}), \end{aligned}$$

which finishes the induction steps.

b) *Step 2: Prove  $\|g^t - p^t\|_{\nabla^2 \psi(p^t)} \leq 2\eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}$ .* By the definition of  $\Phi^t$  and  $\Psi^t$ , we have that

$$\Psi^t(p) = \Phi^t(p) + \eta \langle p, u^t - m^t \rangle.$$

Since  $\nabla \Phi^t(p^t) = 0$  by the first-order optimality and the definition of  $p^t$ , the above relationship between  $\Phi^t$  and  $\Psi^t$  gives that

$$\nabla \Psi^t(p^t) = \eta(u^t - m^t).$$

Since  $g^t = \arg \min(-\Psi^t)$ , and

$$\|\nabla \Psi^t(p^t)\|_{(\nabla^2 \Psi(p^t))^{-1}} = \|\nabla \Psi^t(p^t)\|_{(\nabla^2 \psi(p^t))^{-1}} = \eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}} < \frac{1}{2}$$



by assumption, invoking Lemma A.6 gives

$$\begin{aligned}\|g^t - p^t\|_{\nabla^2\psi(p^t)} &= \|p^t - \arg\min(-\Psi^t)\|_{\nabla^2\psi(p^t)} \\ &\leq 2\|\nabla\Psi^t(p^t)\|_{(\nabla^2\psi(p^t))^{-1}} \\ &= 2\eta\|u^t - m^t\|_{(\nabla^2\psi(p^t))^{-1}}.\end{aligned}\tag{20}$$

c) *Step 3: Lower bound last two terms in (16).*: We first use the inequality in (13) to prove that

$$w(\|p^t - g^t\|_{\nabla^2\psi(p^t)}) + w(\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}) \geq \frac{1}{4} \left( \|p^t - g^t\|_{\nabla^2\psi(p^t)}^2 + \|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2 \right).\tag{21}$$

It suffices to show that  $\|p^t - g^t\|_{\nabla^2\psi(p^t)} \leq 1$  and  $\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})} \leq 1$ .

By (20), we have that

$$\|p^t - g^t\|_{\nabla^2\psi(p^t)} \leq 2\eta\|u^t - m^t\|_{(\nabla^2\psi(p^t))^{-1}} \leq \frac{1}{4},\tag{22}$$

where the last inequality is due to the assumption that  $\eta\|u^t - m^t\|_{(\nabla^2\psi(p^t))^{-1}} \leq \frac{1}{8}$ .

For term  $\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}$ , we need to invoke Lemma A.6 again. Notice that the relationship between  $\Psi^t(p)$  and  $\Phi^t(p)$  is as follows:

$$\Phi^t(p) = \Psi^{t-1}(p) + \eta\langle p, m^t \rangle,$$

which implies

$$\nabla\Phi^t = \nabla\Psi^{t-1} + \eta m^t.$$

Since  $g^{t-1}$  minimize  $\Psi^{t-1}$ , we have  $\nabla\Psi^{t-1}(g^{t-1}) = 0$  by the first-order optimality. Observe that

$$\|\nabla\Phi^t(g^{t-1})\|_{(\nabla^2\Phi(g^{t-1}))^{-1}} = \|\eta m^t\|_{(\nabla^2\psi(g^{t-1}))^{-1}} \leq \frac{1}{2}$$

by assumption. Thus, by Lemma A.6, we have

$$\begin{aligned}\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})} &= \|g^{t-1} - p^t\|_{\nabla^2\psi(g^{t-1})} \\ &\leq 2\|\nabla\Phi^t(g^{t-1})\|_{(\nabla^2\psi(g^{t-1}))^{-1}} \\ &= 2\eta\|m^t\|_{(\nabla^2\psi(g^{t-1}))^{-1}} \leq 1,\end{aligned}\tag{23}$$

where the last inequality is due to the assumption that  $\eta\|m^t\|_{(\nabla^2\psi(g^{t-1}))^{-1}} \leq \frac{1}{2}$ .

Now, we have proven (21). The remaining step is to prove that

$$\sum_{t=1}^T \|p^t - g^t\|_{\nabla^2\psi(p^t)}^2 + \|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2 \geq \frac{1}{4} \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2.$$

By (22) and our assumption, it follows that  $\nabla^2\psi(p^{t-1}) \preceq 2\nabla^2\psi(g^{t-1})$ . Thus, we have

$$\|p^t - g^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 \leq 2\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2.$$

By the triangle inequality for norms, we further have

$$\begin{aligned}\|p^t - p^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 &\leq 2\|p^t - g^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 + 2\|g^{t-1} - p^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 \\ &\leq 4\|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2 + 4\|p^{t-1} - g^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2.\end{aligned}$$

Summing over  $t = 1, \dots, T$  gives the desired results:

$$\begin{aligned}\frac{1}{4} \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 &\leq \sum_{t=1}^T \|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2 + \|p^{t-1} - g^{t-1}\|_{\nabla^2\psi(p^{t-1})}^2 \\ &\leq \sum_{t=1}^T \|p^t - g^{t-1}\|_{\nabla^2\psi(g^{t-1})}^2 + \|p^t - g^t\|_{\nabla^2\psi(p^t)}^2.\end{aligned}$$

□

Now, we revisit the example from [17], in which OFTRL is instantiated with the log-barrier regularizer on the probability simplex. Define the log-barrier regularizer over the probability simplex  $\Delta^{d-1}$  as follows:

$$\psi(p) := - \sum_{a \in [d]} \ln(p(a)).\tag{24}$$

Using this regularizer, and applying Lemma A.9, the authors of [17] show that the OFTRL algorithm admits the following guarantee for any  $p \in \text{relint}(\Delta^{d-1}) := \left\{ p \in \mathbb{R}_{\geq 0}^d : \sum_{a \in [d]} p(a) = 1 \right\}$ :

**Lemma A.10.** *Let  $g^t := \arg \max_{g \in \Omega} \eta \left\langle \sum_{s=1}^t u^s, g \right\rangle - \psi(g)$  be the auxiliary sequence. Assume  $\eta \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}} \leq \frac{1}{8}$  and  $\eta \|m^t\|_{(\nabla^2 \psi(g^{t-1}))^{-1}} \leq \frac{1}{2}$  for all  $t \in [T]$ . Then for any reward sequence  $u^1 \dots u^t \geq 0$  and any  $p \in \text{relint}(\Delta^{d-1})$ , OFTRL with  $\psi$  defined in (24) satisfies*

$$\sum_{t=1}^T \langle p - p^t, u^t \rangle \leq \frac{\psi(p)}{\eta} + 2\eta \sum_{t=1}^T \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}^2 - \frac{1}{16} \eta \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2 \psi(p^{t-1})}^2.$$

*Proof.* We adapt the proof in [17] to the following three steps.

d) *Step 1: Reduce the problem.*: We convert our learning problem to the domain as

$$\Delta^\circ := \left\{ p \in \mathbb{R}_{\geq 0}^{d-1} : \sum_{a \in [d-1]} p(a) \leq 1 \right\},$$

and prove that two problems have the equal regret.

For notational convenience, define the final coordinate as  $p(d) := 1 - \sum_{a \in [d-1]} p(a)$ , so that the full  $d$ -dimensional distribution is implicitly specified. The log-barrier regularizer is then given by:

$$\tilde{\psi}(p) := - \sum_{a \in [d-1]} \ln(p(a)) - \ln \left( 1 - \sum_{a \in [d-1]} p(a) \right).$$

To work in a  $(d-1)$ -dimensional space, we first transform the  $d$ -dimensional reward vector into a  $(d-1)$ -dimensional one. For each  $a \in [d-1]$ , define  $\tilde{u}^t(a) := u^t(a) - u^t(d)$ , and similarly,  $\tilde{m}^t(a) := m^t(a) - m^t(d)$ . This transformation preserves the regret. Specifically, for any  $p, p^t \in \text{int}(\Delta^\circ)$ , we have

$$\begin{aligned} \sum_{a \in [d-1]} \tilde{u}^t(a) (p(a) - p^t(a)) &= \sum_{a \in [d-1]} u^t(a) (p(a) - p^t(a)) - u^t(d) \sum_{a \in [d-1]} (p(a) - p^t(a)) \\ &= \sum_{a \in [d]} u^t(a) (p(a) - p^t(a)), \end{aligned}$$

where the last equality uses the notation  $p(d) := 1 - \sum_{a \in [d-1]} p(a)$ , and similarly for  $\tilde{p}^t(d)$ . Thus, the transformed reward vector yields the same regret expression as the original.

Now, invoking Lemma A.9 (assumptions verified later in Step 4 for clarity of exposition) on  $\tilde{u}^1, \dots, \tilde{u}^T$  gives

$$\begin{aligned} \sum_{t=1}^T \langle p - p^t, u^t \rangle &= \sum_{t=1}^T \langle p - p^t, \tilde{u}^t \rangle \leq \frac{\psi(p)}{\eta} + 2\eta \sum_{t=1}^T \|\tilde{u}^t - \tilde{m}^t\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 \\ &\quad - \frac{1}{4\eta} \sum_{t=1}^T \|p^t - p^{t-1}\|_{\nabla^2 \tilde{\psi}(p^{t-1})}^2. \end{aligned}$$

The lemma follows by showing  $\|\tilde{u}^t - \tilde{m}^t\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 \leq \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}^2$  and  $\|p^t - p^{t-1}\|_{\nabla^2 \tilde{\psi}(p^{t-1})}^2 = \|p^t - p^{t-1}\|_{\nabla^2 \psi(p^{t-1})}^2$ .

e) *Step 2: Prove  $\|p^t - p^{t-1}\|_{\nabla^2 \tilde{\psi}(p^{t-1})}^2 = \|p^t - p^{t-1}\|_{\nabla^2 \psi(p^{t-1})}^2$ .*: Denote by  $\tilde{p} = p^t - p^{t-1}$ . We first show that  $\|\tilde{p}\|_{\nabla^2 \tilde{\psi}(p^{t-1})}^2 = \|\tilde{p}\|_{\nabla^2 \psi(p^{t-1})}^2$ . The Hessian of  $\tilde{\psi}$  can be computed as follows:

$$\nabla^2 \tilde{\psi}(p^{t-1}) = \text{diag} \left( \frac{1}{p^{t-1}(1)^2}, \dots, \frac{1}{p^{t-1}(d-1)^2} \right) + \frac{1}{p^{t-1}(d)^2} \mathbf{1}_{d-1} \mathbf{1}_{d-1}^\top. \quad (25)$$

The Hessian of  $\tilde{\psi}$  can be computed as (25), and the Hessian of  $\psi$  is

$$\nabla^2 \psi(p^{t-1}) = \text{diag} \left( \frac{1}{p^{t-1}(1)^2}, \dots, \frac{1}{p^{t-1}(d)^2} \right).$$

Then, by definition of local norms, we have

$$\|\tilde{p}\|_{\nabla^2 \tilde{\psi}(p^{t-1})}^2 = \sum_{a \in [d-1]} \frac{\tilde{p}(a)^2}{p^{t-1}(a)^2} + \frac{\left(\sum_{a \in [d-1]} \tilde{p}(a)\right)^2}{p^{t-1}(d)^2} = \sum_{a \in [d]} \frac{\tilde{p}(a)^2}{p^{t-1}(a)^2} = \|\tilde{p}\|_{\nabla^2 \psi(p^{t-1})},$$

where the second equality is due to  $\tilde{p}(d) = p^t(d) - p^{t-1}(d) = 1 - \sum_{a \in [d-1]} p^t(a) - 1 + \sum_{a \in [d-1]} p^{t-1}(a) = -\sum_{a \in [d-1]} \tilde{p}(a)$ .

f) *Step 3: Prove  $\|\tilde{u}^t - \tilde{m}^t\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 \leq \|u^t - m^t\|_{(\nabla^2 \psi(p^t))^{-1}}^2$ .* Denote by  $\tilde{u} := \tilde{u}^t - \tilde{m}^t$  and  $u := u^t - m^t$ . Because of our definition  $\tilde{u}^t[a] = u^t(a) - u^t(d)$  and  $\tilde{m}^t[a] = m^t(a) - m^t(d)$  for any  $a \in [d-1]$ , we have the relationship between  $\tilde{u}(a)$  and  $u(a)$  for any  $a \in [d-1]$  as follows.

$$\tilde{u}(a) = \tilde{u}^t(a) - \tilde{m}^t(a) = u^t(a) - m^t(a) - (u^t(d) - m^t(d)) = u(a) - u(d). \quad (26)$$

Now, we show that  $\|\tilde{u}\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 \leq \|u\|_{(\nabla^2 \psi(p^t))^{-1}}^2$ . Since by Sherman–Morrison formula (i.e.,  $(D + \alpha \mathbf{1}\mathbf{1}^\top)^{-1} = D^{-1} - \frac{D^{-1} \mathbf{1}\mathbf{1}^\top D^{-1}}{\frac{1}{\alpha} + \mathbf{1}^\top D^{-1} \mathbf{1}}$ ), we have

$$(\nabla^2 \tilde{\psi}(p^{t-1}))^{-1} = \text{diag}(p^{t-1}(1)^2, \dots, p^{t-1}(d-1)^2) - \frac{(p^{t-1}(p^{t-1})^\top)^2}{\sum_{a \in [d]} p(a)^2},$$

and

$$(\nabla^2 \psi(p^{t-1}))^{-1} = \text{diag}(p^{t-1}(1)^2, \dots, p^{t-1}(d)^2).$$

Then, by the definition of local norms and (26)

$$\begin{aligned} \|\tilde{u}\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 &= \sum_{a \in [d-1]} (u(a) - u(d))^2 p^t(a)^2 - \frac{\left(\sum_{a \in [d-1]} p^t(a)^2 (u(a) - u(d))\right)^2}{\sum_{a \in [d]} p^t(a)^2} \\ &= \sum_{a \in [d]} (u(a) - u(d))^2 p^t(a)^2 - \frac{\left(\sum_{a \in [d]} p^t(a)^2 (u(a) - u(d))\right)^2}{\sum_{a \in [d]} p^t(a)^2} \\ &= \sum_{a \in [d]} (u(a) - u(d))^2 p^t(a)^2 - \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a) - u(d) \sum_{a \in [d]} p^t(a)^2\right)^2}{\sum_{a \in [d]} p^t(a)^2} \\ &= \sum_{a \in [d]} (u(a)^2 - 2u(a)u(d) + u(d)^2) p^t(a)^2 - \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a)\right)^2}{\sum_{a \in [d]} p^t(a)^2} + u(d) \sum_{a \in [d]} 2u(a)p^t(a)^2 - u(d)^2 \sum_{a \in [d]} p^t(a)^2 \\ &= \sum_{a \in [d]} u(a)^2 p^t(a)^2 - \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a)\right)^2}{\sum_{a \in [d]} p^t(a)^2} \\ &= \sum_{a \in [d]} u(a)^2 p^t(a)^2 - 2 \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a)\right)^2}{\sum_{a \in [d]} p^t(a)^2} + \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a)\right)^2}{\sum_{a \in [d]} p^t(a)^2} \\ &= \sum_{a \in [d]} p^t(a)^2 \left( u(a)^2 - 2u(a) \frac{\sum_{a \in [d]} p^t(a)^2 u(a)}{\sum_{a \in [d]} p^t(a)^2} + \frac{\left(\sum_{a \in [d]} p^t(a)^2 u(a)\right)^2}{\left(\sum_{a \in [d]} p^t(a)^2\right)^2} \right) \\ &= \sum_{a \in [d]} p^t(a)^2 \left( u(a) - \frac{\sum_{a \in [d]} p^t(a)^2 u(a)}{\sum_{a \in [d]} p^t(a)^2} \right)^2. \end{aligned}$$

Let  $c_\star = \frac{\sum_{a \in [d]} p(a)^2 u(a)}{\sum_{a \in [d]} p(a)^2}$ , and it is the minimizer of the quadratic function in the last equation. Therefore, we obtain that

$$\|\tilde{u}\|_{(\nabla^2 \tilde{\psi}(p^t))^{-1}}^2 = \|u - c_\star \mathbf{1}\|_{(\nabla^2 \psi(p^t))^{-1}}^2 \leq \|u\|_{(\nabla^2 \psi(p^t))^{-1}}^2.$$

g) *Step 4: Prove the assumptions of Lemma A.9 are satisfied.*: We show below that  $\tilde{\psi}$  satisfies the assumptions of Lemma A.9, i.e., a)  $\nabla^2\psi(p)$  is positive definite; b)  $\tilde{\psi}(p) \geq 0, \forall p \in \text{int}(\Delta^\circ)$  and c)  $\nabla^2\tilde{\psi}(\tilde{p}) \preceq 2\nabla^2\tilde{\psi}(p)$  for any  $p, \tilde{p} \in \text{int}(\Delta^\circ)$  with  $\|p - \tilde{p}\|_{\nabla^2\tilde{\psi}(\tilde{p})} \leq \frac{1}{4}$ .

a) By (25), it is clear that  $\nabla^2\psi(p)$  is positive definite.

b) Since  $p \in \text{int}(\Delta^\circ)$ , we have  $\ln(p(a)) < 0, \forall a \in [d-1]$  and  $1 - \sum_{a \in [d-1]} p(a) \in [0, 1]$ , so that  $\ln\left(1 - \sum_{a \in [d-1]} p(a)\right) \leq 0$ .

Thus,  $\tilde{\psi}(p) \geq 0, \forall p \in \text{int}(\Delta^\circ)$ .

c) Since  $\|p - \tilde{p}\|_{\nabla^2\tilde{\psi}(\tilde{p})} \leq \frac{1}{4}$ , by invoking Lemma A.4, we have that

$$\nabla^2\tilde{\psi}(\tilde{p}) \preceq \frac{1}{(1 - \|\tilde{p} - p\|_{\nabla^2\tilde{\psi}(p)})^2} \nabla^2\tilde{\psi}(p) \leq \frac{16}{9} \nabla^2\tilde{\psi}(p) \leq 2\nabla^2\tilde{\psi}(p).$$

□

### C. Proof of Lemma V.1

*Proof.* The proof follows the two steps below.

1) We claim that for any  $t \in [T]$ :

$$\begin{aligned} & \sum_{a \in \mathcal{A}_n} \|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\|_{*,q_a^t}^2 \\ & \leq 2 \sum_{a \in \mathcal{A}_n} |u_n^t(a) - m_n^t(a)| \mathbf{1}[a_n^t = a] + 2\|p_n^t - p_n^{t-1}\|_1. \end{aligned} \quad (27)$$

2) Next, we complete the proof by demonstrating that when summed over  $t \in [T]$ , we have

$$\sum_{t=1}^T \sum_{a \in \mathcal{A}_n} |u_n^t(a) - m_n^t(a)| \mathbf{1}[a_n^t = a] = \sum_{t=1}^T \|u_n^t - u_n^{t-1}\|_1.$$

a) *Step 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

$$\begin{aligned} & \|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\|_{*,q_a^t}^2 \\ & = \sum_{a' \in \mathcal{A}_n} (q_a^t(a'))^2 (p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t + p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t)^2 \\ & \leq 2 \sum_{a' \in \mathcal{A}_n} (q_a^t(a'))^2 \left( (p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t)^2 + (p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t)^2 \right) \\ & = 2\|p_n^t(a)\hat{u}_n^t - p_n^t(a)m_n^t\|_{*,q_a^t}^2 + 2\|p_n^t(a)m_n^t - p_n^{t-1}(a)m_n^t\|_{*,q_a^t}^2, \end{aligned} \quad (28)$$

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

Summing over  $a \in \mathcal{A}_n$ , the first term on the RHS of (28) can be further bounded as follows.

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

where the first inequality is due to  $\frac{p_n^t(a)q_a^t(a')}{p_n^t(a')} \leq 1$  and  $|u_n^t(a') - m_n^t(a')| \leq 1$ . The last equality 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 (28), we have that

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

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

b) *Step 2.*: Then, Step 2 is obtained by summing over  $t \in [T]$  as follows.

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

□

#### D. Proof of Lemma A.11

**Lemma A.11.** When  $\eta \leq \frac{1}{16}$ , we have that

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

Before we prove Lemma A.11, we need the following lemma.

**Lemma A.12** (Multiplicative Stability for Bandits). For  $\eta \leq \frac{1}{16}$ ,  $\|u_n^t\|_\infty \leq 1$  and  $\|p_n^{t-1}(a)m_n^t\|_\infty \leq 1$  for all  $t \in [T]$  and  $a \in \mathcal{A}_n$ , we have that with the log-barrier regularizer:

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

*Proof of Lemma A.12.* By the triangle inequality, we have

$$\sum_{a \in \mathcal{A}_n} \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}} \leq \sum_{a \in \mathcal{A}_n} \|q_a^t - g_a^{t-1}\|_{q_a^{t-1}} + \sum_{a \in \mathcal{A}_n} \|g_a^{t-1} - q_a^{t-1}\|_{q_a^{t-1}}.$$

Then, we need to show that both terms on the RHS of the above inequality are no larger than  $\frac{1}{4}$  to prove the lemma. Let  $\Psi_a^t(g) := \eta \left\langle \sum_{s=1}^t p_n^s(a) \hat{u}_n^s, g \right\rangle - \psi(g)$ , and denote by  $g_a^t := \arg \max_{g \in \text{relint}(\Delta(\mathcal{A}_n))} \Psi_a^t(g)$  the maximizer of  $\Psi_a^t(g)$ . Similarly, let

$\Phi_a^t(q) := \eta \left\langle p_n^{t-1}(a)m_n^t + \sum_{s=1}^{t-1} p_n^s(a) \hat{u}_n^s, q \right\rangle - \psi(q)$ , and denote by  $q_a^t := \arg \max_{q \in \text{relint}(\Delta(\mathcal{A}_n))} \Phi_a^t(q)$  be the maximizer of  $\Phi_a^t(q)$ .

a) *Step 1: Bound  $\sum_{a \in \mathcal{A}_n} \|q_a^t - g_a^{t-1}\|_{q_a^{t-1}}$ .* We can apply the same reasoning as in (23). Notice that

$$\Phi_a^t(g) = \Psi_a^{t-1}(g) + \eta \langle p_n^{t-1}(a)m_n^t, g \rangle,$$

which implies

$$\nabla \Phi_a^t = \nabla \Psi_a^{t-1} + \eta p_n^{t-1}(a)m_n^t.$$

Since  $g_a^{t-1}$  minimizes  $\Psi_a^{t-1}$ , we have  $\nabla \Psi_a^{t-1}(g_a^{t-1}) = 0$  by the first-order optimality. Thus, we have

$$\|\nabla \Phi_a^t(g_a^{t-1})\|_{*,g_a^{t-1}} = \eta \|p_n^{t-1}(a)m_n^t\|_{*,g_a^{t-1}} \leq \eta \leq \frac{1}{2},$$

where the first inequality is due to  $g_a^{t-1} \in \text{relint}(\Delta(\mathcal{A}_n))$ ,  $p_n^{t-1}(a) \leq 1$  and  $m_n^t \leq 1$ . Therefore, we can invoke Lemma A.6 to have

$$\begin{aligned} \sum_{a \in \mathcal{A}_n} \|q_a^t - g_a^{t-1}\|_{q_a^{t-1}} &= \sum_{a \in \mathcal{A}_n} \|g_a^{t-1} - \arg \min_{q \in \text{relint}(\Delta(\mathcal{A}_n))} (-\Phi^t(q))\|_{q_a^{t-1}} \\ &\leq 2 \sum_{a \in \mathcal{A}_n} \|\nabla \Phi_a^t(g_a^{t-1})\|_{*, g_a^{t-1}} \\ &= 2\eta \sum_{a \in \mathcal{A}_n} \|p_n^{t-1}(a)m_n^t\|_{*, g_a^{t-1}} \\ &\leq 2\eta \|m_n^t\|_{*, g_a^{t-1}} \leq \frac{1}{8} < \frac{1}{4}, \end{aligned}$$

where the second inequality is by the assumption that  $\eta \leq \frac{1}{16}$ .

b) *Step 2: Bound  $\sum_{a \in \mathcal{A}_n} \|g_a^{t-1} - q_a^{t-1}\|_{q_a^{t-1}}$ .* Similar to (20), we first observe that

$$\Psi_a^{t-1}(q) = \Phi_a^{t-1}(q) + \eta \langle q, p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1} \rangle,$$

which implies

$$\nabla \Psi_a^{t-1} = \nabla \Phi_a^{t-1} + \eta (p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}).$$

Since  $q_a^{t-1}$  minimizes  $\Phi_a^{t-1}$ , we have  $\nabla \Phi_a^{t-1}(q_a^{t-1}) = 0$  by the first-order optimality. Thus, we have

$$\nabla \Psi_a^{t-1}(q_a^{t-1}) = \eta (p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}).$$

Then, we want to invoke Lemma A.6 again, and we need to show  $\|\nabla \Psi_a^{t-1}(q_a^{t-1})\|_{*, q_a^{t-1}} \leq \frac{1}{2}$ . By the triangle inequality, we have that

$$\begin{aligned} \|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\|_{*, q_a^{t-1}} &\leq \|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-1}(a)m_n^{t-1}\|_{*, q_a^{t-1}} \\ &\quad + \|p_n^{t-1}(a)m_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\|_{*, q_a^{t-1}}. \end{aligned} \quad (29)$$

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

$$\begin{aligned} &\|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-1}(a)m_n^{t-1}\|_{*, q_a^{t-1}} \\ &\leq \sqrt{\sum_{a' \in \mathcal{A}_n} (q_a^{t-1}(a'))^2 \left( \frac{p_n^{t-1}(a)\mathbf{1}[a_n^{t-1} = a'](u_n^t(a') - m_n^t(a'))}{p_n^{t-1}(a')} \right)^2} \\ &\leq \sqrt{\sum_{a' \in \mathcal{A}_n} \left( \frac{p_n^{t-1}(a)q_a^{t-1}(a')}{p_n^{t-1}(a')} \right)^2 \mathbf{1}[a_n^{t-1} = a']} = \frac{p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})}, \end{aligned} \quad (30)$$

where the second inequality is due to that  $|u_n^t(a) - m_n^t(a)| \leq 1$ . The second term on the RHS of (29) can be bounded as follows

$$\begin{aligned} &\|p_n^{t-1}(a)m_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\|_{*, q_a^{t-1}} \\ &= \sqrt{\sum_{a' \in \mathcal{A}_n} (q_a^{t-1}(a'))^2 (p_n^{t-1}(a)m_n^{t-1}(a') - p_n^{t-2}(a)m_n^{t-1}(a'))^2} \\ &\leq \sqrt{\sum_{a' \in \mathcal{A}_n} q_a^{t-1}(a') (p_n^{t-1}(a) - p_n^{t-2}(a))^2} \\ &= p_n^{t-1}(a) - p_n^{t-2}(a). \end{aligned} \quad (31)$$

Then, substituting (30) and (31) into (29) gives

$$\begin{aligned} \|\nabla \Psi^{t-1}(q_a^{t-1})\|_{*, q_a^{t-1}} &= \eta \|p_n^{t-1}(a)\hat{u}_n^{t-1} - p_n^{t-2}(a)m_n^{t-1}\|_{*, q_a^{t-1}} \\ &\leq \eta \frac{p_n^{t-1}(a)q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})} + \eta (p_n^{t-1}(a) - p_n^{t-2}(a)) \\ &\leq 2\eta \leq \frac{1}{8} < \frac{1}{2}, \end{aligned} \quad (32)$$

where the first inequality is due to that  $p_n^{t-1}(a') = \sum_{a \in \mathcal{A}_n} p_n^{t-1}(a) q_a^{t-1}(a')$ , and the last inequality is due to the assumption that  $\eta \leq \frac{1}{16}$ . Then, by invoking Lemma A.6, we have that

$$\begin{aligned}
\sum_{a \in \mathcal{A}_n} \|g_a^{t-1} - q_a^{t-1}\|_{q_a^{t-1}} &= \sum_{a \in \mathcal{A}_n} \|q_a^{t-1} - \arg \min_{g \in \text{relint}(\Delta(\mathcal{A}_n))} (-\Psi^{t-1}(g))\|_{q_a^{t-1}} \\
&\leq 2 \sum_{a \in \mathcal{A}_n} \|\nabla \Psi^{t-1}(q_a^{t-1})\|_{*, q_a^{t-1}} \\
&\leq 2 \sum_{a \in \mathcal{A}_n} \eta \|p_n^{t-1}(a) \hat{u}^{t-1} - p_n^{t-1}(a) m_n^{t-2}\|_{*, q_a^{t-1}} \\
&\leq 2 \sum_{a \in \mathcal{A}_n} \eta \frac{p_n^{t-1}(a) q_a^{t-1}(a_n^{t-1})}{p_n^{t-1}(a_n^{t-1})} + 2\eta \sum_{a \in \mathcal{A}_n} (p_n^{t-1}(a) - p_n^{t-2}(a)) \\
&\leq 2\eta + 0 \leq \frac{1}{8} < \frac{1}{4}.
\end{aligned}$$

□

Then, we are ready to give the proof of Lemma A.11 as follows.

*Proof of Lemma A.11.* 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

$$(\mu_a^t)^2 \leq \sum_{a' \in \mathcal{A}_n} \left( 1 - \frac{q_a^t(a')}{q_a^{t-1}(a')} \right)^2 = \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2. \quad (33)$$

According to Lemma A.12, we have that  $\sum_{a \in \mathcal{A}_n} \mu_a^t \leq 0.5$ . By the definition of  $\mu_a^t$ , 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.7, 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

$$\begin{aligned}
\prod_{u \in \mathcal{A}_n} (1 - \mu_u^t) \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^{t-1}(v) &\leq \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^t(v) \\
&\leq \prod_{u \in \mathcal{A}_n} (1 + \mu_u^t) \prod_{(u,v) \in E(\mathcal{T}_a)} q_u^{t-1}(v).
\end{aligned} \quad (34)$$

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 (34) 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.12.

Since  $\Sigma^t = \sum_{a \in \mathcal{A}_n} \Sigma_a^t$ , we have the lower and upper bounds for  $\Sigma^t$  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 (33), we have that

$$\begin{aligned}
\|p_n^t - p_n^{t-1}\|_1^2 &\leq 64 \left( \sum_{a \in \mathcal{A}_n} \mu_a^t \right)^2 \leq 64 A_n \sum_{a \in \mathcal{A}_n} (\mu_a^t)^2 \\
&\leq 64 A_n \sum_{a \in \mathcal{A}_n} \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2.
\end{aligned} \tag{35}$$

□

#### E. Proof of Theorem V.2

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

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



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:

$$\begin{aligned} R_n^{\text{swa}}(T) &= \max_{F \in \mathcal{F}_n} \mathbf{E} \left[ \sum_{t=1}^T \sum_{a \in \mathcal{A}_n} \mathbf{1}[a_n^t = a] (u_n^t(F(a)) - u_n^t(a)) \right] \\ &= \max_{F \in \mathcal{F}_n} \mathbf{E} \left[ \sum_{t=1}^T \mathbf{E} \left[ \sum_{a \in \mathcal{A}_n} p_n^t(a) (u_n^t(F(a)) - u_n^t(a)) \mid \mathcal{F}_{t-1} \right] \right] \\ &= \max_{F \in \mathcal{F}_n} \sum_{t=1}^T \sum_{a \in \mathcal{A}_n} \mathbf{E} [p_n^t(a) \hat{u}_n^t(F(a)) - \langle q_a^t, p_n^t(a) \hat{u}_n^t \rangle]. \end{aligned} \quad (36)$$

**Step 2:** Notice that for any  $F \in \mathcal{F}_n$ :

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

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

Since Lemma A.10 requires that  $q \in \text{relint}(\Delta(\mathcal{A}_n))$ , but  $q_a^* \in \Delta(\mathcal{A}_n)$ . Let  $q_a^* := \arg \max_{q \in \Delta(\mathcal{A}_n)} \mathbf{E}[R_a^T(q)]$ . We get around with this by letting  $\tilde{q}_a := (1 - \frac{1}{T}) q_a^* + \frac{1}{T} q_a^c$ , where  $q_a^c(a') = \frac{1}{A_n}, \forall a' \in \mathcal{A}_n$ , and thus  $\tilde{q}_a \in \text{relint}(\Delta(\mathcal{A}_n))$ . Now we can write  $R_a^T$  as follows.

$$\begin{aligned} \mathbf{E} [R_a^T(q_a^*)] &= \mathbf{E} \left[ \sum_{t=1}^T \langle q_a^* - \tilde{q}_a, p_n^t(a) \hat{u}_n^t \rangle + \sum_{t=1}^T \langle \tilde{q}_a - q_a^t, p_n^t(a) \hat{u}_n^t \rangle \right] \\ &= \mathbf{E} \left[ \sum_{t=1}^T \frac{1}{T} \langle q_a^* - q_a^c, p_n^t(a) \hat{u}_n^t \rangle \right] + \mathbf{E} [R_a^T(\tilde{q}_a)]. \end{aligned} \quad (37)$$

We can bound the first term of the RHS in the above equation as follows.

$$\begin{aligned} \mathbf{E} \left[ \sum_{t=1}^T \frac{1}{T} \langle q_a^* - q_a^c, p_n^t(a) \hat{u}_n^t \rangle \right] &= \mathbf{E} \left[ \sum_{t=1}^T \frac{1}{T} \langle q_a^* - q_a^c, p_n^t(a) u_n^t \rangle \right] \\ &\leq \mathbf{E} \left[ \sum_{t=1}^T \frac{1}{T} \|q_a^* - q_a^c\|_1 \|p_n^t(a) u_n^t\|_\infty \right] \leq \mathbf{E} \left[ \sum_{t=1}^T \frac{2}{T} p_n^t(a) \right], \end{aligned}$$

where the last inequality is due to that  $\|q_a^* - q_a^c\|_1 \leq \|q_a^*\|_1 + \|q_a^c\|_1 \leq 2$  and  $u_n^t \leq 1$ .

Next, we can invoke Lemma A.10 to bound  $R_a^T(\tilde{q}_a)$ , which requires  $\eta \|p_n^t(a) \hat{u}_n^t - p_n^{t-1}(a) m_n^t\|_{*, q_a^t} \leq \frac{1}{8}$  and  $\eta \|p_n^{t-1}(a) m_n^t\|_{*, g_a^{t-1}} \leq \frac{1}{2}$ , where  $g_a^t \in \text{relint}(\Delta(\mathcal{A}_n))$  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 Lemma A.10 is satisfied when  $\eta \leq \frac{1}{16}$ . According to (29) to (32), we have that

$$\begin{aligned} &\eta \|p_n^t(a) \hat{u}_n^t - p_n^{t-1}(a) m_n^t\|_{*, q_a^t} \\ &\leq \eta \left( \|p_n^t(a) \hat{u}_n^t - p_n^t(a) m_n^t\|_{*, q_a^t} + \|p_n^t(a) m_n^t - p_n^{t-1}(a) m_n^t\|_{*, q_a^t} \right) \\ &\leq \eta \left( \frac{p_n^t(a) q_a^t(a_n^t)}{p_n^t(a_n^t)} + p_n^t(a) - p_n^{t-1}(a) \right) \leq 2\eta \leq \frac{1}{8}, \end{aligned} \quad (38)$$

where the third inequality is due to  $\frac{p_n^t(a) q_a^t(a_n^t)}{p_n^t(a_n^t)} \leq 1$  and  $p_n^t(a) - p_n^{t-1}(a) \leq 1$ , and the last inequality is due to  $\eta \leq \frac{1}{16}$ .

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

$$\eta \|p_n^{t-1}(a)m_n^t\|_{*,q} \leq \eta \|p_n^{t-1}(a)m_n^t\|_\infty \leq \eta \leq \frac{1}{2}.$$

Then, we can invoke Lemma A.10 to bound  $R_a^T(\tilde{q}_a)$  as follows:

$$R_a^T(\tilde{q}_a) \leq \frac{\psi(\tilde{q}_a)}{\eta} + 2\eta \sum_{t=1}^T \|p_n^t(a)\hat{u}_n^t - p_n^{t-1}(a)m_n^t\|_{*,q_a^t}^2 - \frac{1}{16\eta} \sum_{t=1}^T \|q_a^t - q_a^{t-1}\|_{q_a^{t-1}}^2. \quad (39)$$

Now, we give a bound of  $\psi(\tilde{q}_a)$  by invoking Lemma A.2 as follows.

$$\psi(\tilde{q}_a) \leq \psi(q_a^c) + A_n \ln T = A_n \ln A_n + A_n \ln T \leq 2A_n \ln T, \quad (40)$$

where we use the fact that  $\pi(\tilde{q}_a; q_a^c) \leq 1 - \frac{1}{T}$  by the definition of  $\tilde{q}_a$  (see definition of  $\pi(\cdot; \cdot)$  in (9)).

Now, substituting (39) and (40) into (37) gives:

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

**Step 3:** Summing over  $a \in \mathcal{A}_n$  and invoking Lemmas V.1 and A.11, we have the swap regret bounded as follows:

$$\begin{aligned} \sum_{a \in \mathcal{A}_n} \mathbf{E}[R_a^T(q^*)] &\leq \mathbf{E} \left[ 2 + \frac{2(A_n)^2 \ln T}{\eta} + 4\eta \sum_{t=1}^T \|u_n^t - u_n^{t-1}\|_1 + 4\eta \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1 \right. \\ &\quad \left. - \frac{1}{1024A_n\eta} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2 \right]. \end{aligned}$$

Next, for any  $t \in [T]$  and  $a \in \mathcal{A}_n$ , since

$$u_n^t(a) = \mathbf{E}_{a_{-n} \sim p_{-n}^t} [u_n(a; a_{-n})] = \sum_{a_{-n}} p_{-n}^t(a_{-n}) u_n(a; a_{-n}),$$

we have that

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

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:

$$\begin{aligned} R_n^{\text{swa}}(T) &\leq \mathbf{E} \left[ 2 + \frac{2(A_n)^2 \ln T}{\eta} + 4\eta \sum_{t=1}^T \sum_{m \in [N]} \|p_m^t - p_m^{t-1}\|_1 \right. \\ &\quad \left. - \frac{1}{1024A_n\eta} \sum_{t=1}^T \|p_n^t - p_n^{t-1}\|_1^2 \right]. \end{aligned}$$

□