A Near-optimal High-probability Swap-regret Upper Bound for Multi-agent Bandits in Unknown General-sum Games

Zhiming Huang[1] **Jianping Pan**[1]

[1]Department of Computer Science, University of Victoria, BC, Canada

Abstract

In this paper, we study a multi-agent bandit problem in an unknown general-sum game repeated for a number of rounds (i.e., learning in a black-box game with bandit feedback), where a set of agents have no information about the underlying game structure and cannot observe each other's actions and rewards. In each round, each agent needs to play an arm (i.e., action) from a (possibly different) arm set (i.e., action set), and *only* receives the reward of the *played* arm that is affected by other agents' actions. The objective of each agent is to minimize her own cumulative swap regret, where the swap regret is a generic performance measure for online learning algorithms. We are the first to give a near-optimal high-probability swap-regret upper bound based on a refined martingale analysis for the exponential-weighting-based algorithms with the implicit exploration technique, which can further bound the expected swap regret instead of the pseudo-regret studied in the literature. It is also guaranteed that correlated equilibria can be achieved in a polynomial number of rounds if the algorithm is played by all agents. Furthermore, we conduct numerical experiments to verify the performance of the studied algorithm.

1 INTRODUCTION

The *multi-armed bandit (MAB)* is a theoretical model for online learning problems. The name comes from imagining a gambler needs to play one of the arms on a slot machine in each round. If an arm is played, then the gambler will receive a random reward. The objective of the gambler is to accumulate as many rewards as possible within T rounds. As the information about which arm can return the highest rewards is not a prior knowledge, the gambler faces a dilemma in each round between playing the currently best arm (i.e., exploitation) or playing other arms to learn more about their rewards (i.e., exploration).

To adapt to more complex scenarios in reality, many variants of MABs have been proposed. In this paper, we study a variant called *multi-agent bandits in an unknown general-sum game (MAB-UG)*, motivated by many real-world problems such as end-to-end congestion control in computer networks. In this case, each host has no information about others and needs to choose a transmission rate, hoping to maximize its throughput without congesting the network. Another example is the medium access control in wireless communications, where a set of devices need to access a shared communication channel to send packets in each time slot.

The MAB-UG setting can be referred to as the black-box game studied in Nax et al. [2016], where a set of agents $\mathcal{N}$, each associated with K_n (possibly different) arms (i.e., actions), are playing an unknown general-sum game repeated for T rounds. All agents have no information about the structure of the underlying game and cannot observe each other's actions and rewards. In each round, each agent needs to play an arm a_n^t from an arm set A_n, and observes the corresponding reward/loss.

The only information is the observed reward/loss for their own played arm in each round. Thus, each agent is facing a non-stochastic multi-armed bandit problem with adaptive (i.e., non-oblivious) adversaries. The objective for each agent is to accumulate as many rewards as possible and the empirical joint distribution of all agents' actions reaches an ϵ-correlated equilibrium [Aumann, 1974], a concept more general than the well-known Nash equilibrium, within T rounds. Intuitively, the ϵ-correlated equilibrium is a state that the expected incentives (e.g., the reward difference) for each agent to deviate from a suggested action are no more than $\epsilon \geq 0$, where the expectation is taken with respect to the joint distribution of all agents' actions.

As each agent has very limited knowledge about the environment and can only learn from the rewards of the played

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

arm that is affected by others' actions in each round, the algorithm to address MAB-UG must be carefully designed to balance the tradeoff between exploration and exploitation. The performance of an algorithm is usually measured by *regret*. The most oft-used definition of regret in the bandit literature is called the *external regret* [Cesa-Bianchi and Lugosi, 2006, Lattimore and Szepesvári, 2020], which measures the performance loss of an algorithm against a set of competitors always playing a fixed action. However, minimizing the external regret is not enough for MAB-UG, as another objective is to achieve the ϵ-correlated equilibrium. Fortunately, it is proved in Hart and Mas-Colell [2000], Cesa-Bianchi and Lugosi [2006] that if every agent plays an algorithm that minimizes *internal regret*, then the empirical joint distribution of actions converges to an ϵ-correlated equilibrium. The internal regret is defined to be the performance loss for an algorithm that plays arm a instead of playing another arm a'. In this paper, we study a stronger regret notion called *swap regret* introduced by Blum and Mansour [2007], which is a generalization of the above two regrets, comparing the performance of a learning algorithm against a larger set of competitors. The swap regret uses swap functions F that take the arms played by an algorithm as input and output the arms to be compared. Thus, by changing the swap functions, the swap regret can boil down to external regret and internal regret.

The swap regret has been extensively studied in terms of *pseudo-regret* (or *weak regret*) [Blum and Mansour, 2007, Stoltz, 2005, Ito, 2020], i.e., $\max_F \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t = a]r_{a,F(a)}\right]$, where $r_{a,F(a)}$ is the instantaneous swap regret with arm a and swap function F, and *conditionally expected swap regret* [Jin et al., 2022], i.e., $\max_F \sum_{t=1}^{T}\sum_{a\in A_n} p_a^t r_{a,F(a)}$, which also bounds the pseudo-regret by taking expectation on the randomness of algorithms. However, bounding the above regret can only guarantee the expected swap regret (i.e., $\mathbf{E}\left[\max_F \sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t = a]r_{a,F(a)}\right]$) is bounded when adversaries are not adaptive (i.e., oblivious) [Audibert and Bubeck, 2010], but each agent in MAB-UG is facing other agents as adaptive adversaries. Thus, a more meaningful but challenging bound is on the instantaneous swap regret (i.e., $\max_F \sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t = a]r_{a,F(a)}$) for any sequence of actions and rewards, which is helpful not only to equilibrium convergence but also to the bound of the expected swap regret with respect to all agents' randomness.

The main contribution of our work is to give an instantaneous swap regret analysis for the exponential-weighting-based algorithms called *learning for correlated equilibrium with implicit exploration (LCE-IX)*. LCE-IX is based on the swap-regret-minimizing framework proposed by Blum and Mansour [2007], and the main idea is to call K_n exponential-weighting-based algorithms with the Implicit eXploration (IX) technique [Kocák et al., 2014, Neu, 2015] as subroutines. Then, the probability of selecting an arm is obtained by the Markov steady-state distribution of the Markov process among K_n subroutines, and the reward/loss is proportionally fed to the subroutines for updates.

However, the existing concentration inequality for the IX technique cannot be simply applied to the analysis of swap regret. The main difficulties are twofold. First, the swap regret is only equivalent to the sum of the external regret for subroutine algorithms in expectation. In this sense, the existing concentration inequality for IX can only give a high-probability bound on the conditionally expected swap regret. When we analyze the instantaneous swap regret, we cannot convert it directly to the sum of the instantaneous external regret for subroutine algorithms. Second, the reward/loss of each arm is a result of all agents' actions, which is not determined at the beginning of each round as in the single-agent bandit setting (see more discussions in Sec. 3).

To address this problem, we prove a novel general-form concentration inequality between the IX loss estimator and the swapped loss based on a refined martingale analysis by treating the K_n subroutine algorithms as a whole. Based on this concentration inequality, we show that with probability at least $1-\delta$ for $\delta \in (0,1)$, the instantaneous swap regret is bounded in $O(K_n\sqrt{T\log(K_n/\delta)})$ for each agent $n \in \mathcal{N}$.

Furthermore, by integrating the tails of this high-probability bound for the instantaneous swap regret, we show the expected swap-regret bound is in $O(K_n\sqrt{T\log(K_n)})$ with respect to all agents' randomness. The above swap-regret bounds are near-optimal with an $O(\sqrt{K_n})$ gap from the swap-regret lower bound by Ito [2020] for a related model. It is also guaranteed that LCE-IX can converge to ϵ-correlated equilibria for unknown general-sum games in a polynomial number of rounds if the algorithm is played by all agents. Numerical experiments verifies the performance of LCE-IX.

The rest of the paper is organized as follows. In Sec. 2, we review the works that are most related to MAB-UG. The problem settings are described in Sec. 3. The LCE-IX algorithm is proposed in Sec. 4, with analytical results presented in Sec. 5. The experiment results are shown, compared and discussed in Sec. 6. Sec. 7 concludes the paper. The detailed proofs of the swap-regret upper bound are deferred to the Appendix in the supplementary materials.

2 RELATED WORKS

Multi-agent bandits: Multi-agent bandits consider a group of agents participating in decision making, and aim to improve learning efficiency through collaborations. The works about multi-agent bandits are mainly focused on improving rewards by communication [Buccapatnam et al., 2015, Chakraborty et al., 2017, Kolla et al., 2018, Vial et al., 2021], identifying the best arm to avoid collision [Bubeck et al., 2020, Liu and Zhao, 2010, Hillel et al., 2013, Szorenyi et al., 2013, Jamieson and Nowak, 2014], and voting for playing arms [Dubey et al., 2020]. All the above bandit settings assume the arm set for each agent is identical, and the reward for an agent does not depend on the actions of other agents, or just follows a simple collision model. MAB-UG considers (possibly) varied arm sets for different agents and more general competitions among agents.

Learning in games: The history of learning in games can be traced back to the fictitious play for the two-player zero-sum games [Brown, 1949, Robinson, 1951]. Nevertheless, such a fictitious play requires that the decisions of opponents can be observed, and thus it cannot be applied to the unknown games where the agents can only observe their own outcomes (or rewards). To address the challenges of unknown games, online learning has been introduced by many works for specific games such as potential games [Coucheney et al., 2015, Cohen et al., 2017, Bielawski et al., 2021, Mguni et al., 2021], and mean-field games [Min and Hu, 2021, Wang et al., 2021, Xie et al., 2021]. However, the above solutions for specific games depend on corresponding properties (e.g., potential functions for potential games), and thus cannot be easily extended to the general-sum games. Thus, we focus on the learning in the unknown general-sum games (i.e., black-box games [Nax et al., 2016]), which is a basic case of learning in general-sum Markov games [Littman, 1994].

Regarding the unknown general-sum games, there are mainly two lines of research depending on the observability of rewards. If the reward of an action can be observed regardless of whether it is played or not, we call it the *full-information* model [Cesa-Bianchi and Lugosi, 2006], and if only the reward of a played action can be observed, then it is the *partial-information* model (i.e., *bandit* feedback). Recent years have witnessed steady progress in learning general-sum games in the full-information model [Krichene et al., 2015, Palaiopanos et al., 2017, Chen and Peng, 2020, Daskalakis et al., 2021, Anagnostides et al., 2022, Farina et al., 2022]. However, the results for the full-information model cannot be easily extended to the partial-information model, as less information is observed in each round, which makes the partial information model more challenging. The first work that addressed the unknown general-sum games with bandit feedback is Auer et al. [2002], where an exponential-weighting-based technique is proposed to minimize the external regret. However, it is typically one of the goals in general games to search for correlated equilibria, and it is shown in Cesa-Bianchi and Lugosi [2006] that only minimizing external regret cannot achieve this goal.

Table 1: Swap-regret bounds for *exponential-weighting*-based algorithms with bandit feedback

Upper bound, Computational cost, Regret notion	Lower bound
$O\left(\sqrt{TK_n^3\log(K_n)}\right)$, poly-time, pseudo-regret [Blum and Mansour, 2007]	$\Omega\left(\sqrt{TK_n}\right)$ [Blum and Mansour, 2007]
$O\left(\sqrt{TK_n^2\log(K_n)}\right)$, exp-time, pseudo-regret [Stoltz, 2005]	
$O\left(\sqrt{TK_n^2\log(K_n)}\right)$, poly-time, pseudo-regret [Ito, 2020]	$\Omega\left(\sqrt{TK_n\log(K_n)}\right)$ [Ito, 2020]
$O\left(\sqrt{TK_n^2\log(K_n/\delta)}\right)$, poly-time, conditionally expected regret [Jin et al., 2022]	
$O\left(\sqrt{TK_n^2\log(K_n/\delta)}\right)$, poly-time, instantaneous regret (our work, Theorem 5.3)	
$O\left(\sqrt{TK_n^2\log(K_n)}\right)$, poly-time, expected regret (our work, Theorem 5.4)	

Blum and Mansour [2007] generalized the notion of external and internal regrets to the swap regret, and proposed a polynomial-time swap-regret-minimizing framework based on K_n external-regret-minimizing subalgorithms, where K_n is the number of arms. They proved that if the external pseudo-regret of each subalgorithm can be represented by a concave function $r(T)$, where T is the time horizon, and if the dependency on K_n is ignored in $r(T)$, the swap pseudo-regret of their proposed algorithm is $K_n \cdot r(T)$. Therefore, as each exponential-weighting subalgorithm has an external-pseudo-regret bound of $r(T) = O(\sqrt{TK_n\log(K_n)})$ [Auer et al., 2002], the analysis of Blum and Mansour [2007] gives a pseudo-regret bound of $O(K_n\sqrt{TK_n\log(K_n)})$ for their proposed algorithm.

Later, this bound was improved by Stoltz [2005] to $O(K_n\sqrt{T\log(K_n)})$ but with an exponential computation complexity. On the other hand, Ito [2020] improved the upper bound for the swap pseudo-regret to $K_n \cdot r(T/K_n)$ with a polynomial-time algorithm by adding another layer of randomness to the original framework [Blum and Mansour, 2007], where in each round only one subroutine is selected according to the calculated Markov steady distribution. The selected subroutine selects an arm, and the reward will be *entirely* fed to this subroutine algorithm for updates. The modified framework gives a pseudo-regret of $O(\sqrt{TK_n^2\log(K_n)})$ for the exponential-weighting-based subalgorithms.[1] It was also proved by Ito [2020] that the lower bound for swap regret is $\Omega\left(\sqrt{TK_n\log(K_n)}\right)$, which is tight in the full-information but not partial-information models. Recently, Jin et al. [2022] proved a high-probability bound of $O(K_n\sqrt{T\log(K_n/\delta)})$ for the conditionally expected swap regret, which can bound the pseudo-regret by integrating the tails. Table 1 gives a summary of the swap-regret bounds for exponential-weighting-based algorithms with bandit feedback.

Thus, we are the first to prove a high-probability bound of $O(K_n\sqrt{T\log(K_n/\delta)})$ for the instantaneous swap regret and bound the expected swap regret in $O(K_n\sqrt{T\log(K_n)})$ with respect to all agents' randomness, which is near-optimal because of an $O(\sqrt{K_n})$ gap from the swap-regret lower bound [Ito, 2020] despite for full-information models.

[1] The swap regret for the modified framework by Ito [2020] can be tighter if mirror descent algorithms [Zimmert and Seldin, 2019] are used. However, in this paper, we only discuss the swap-regret bound for the exponential-weighting-based algorithms.

3 PROBLEM FORMULATION

3.1 THE MAB-UG MODEL

In MAB-UG, the reward of each agent's action will be affected by the actions of other agents, and each agent has no prior knowledge about the environment such as the number of agents, the reward of each action, and the actions of other agents. A simple example of MAB-UG with two agents and two arms for each agent is shown in Fig. 1, where in the current round, Agent 1 plays arm a and only observes a normalized reward of 0.8, and Agent 2 plays arm c and only observes a normalized reward of 0.2. Both agents have no information about the arm played by the other agent, nor the rewards of the arms that are not played.

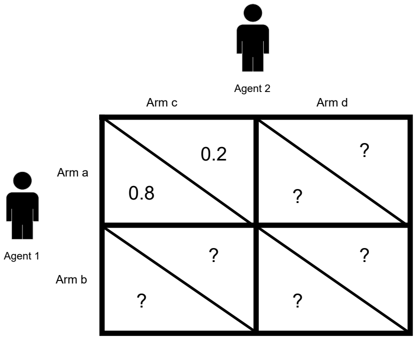


Figure 1: An example of MAB-UG with two agents and two arms for each agent.

Formally, let $\mathcal{N} := \{1, \ldots, N\}$ be the set of all agents and each agent $n \in \mathcal{N}$ is associated with a finite set of arms (i.e., actions) A_n with size K_n. The arm set for each agent is not required to be identical. Let $\mathcal{A} := \prod_{n\in\mathcal{N}} A_n$ be the space of all such arm sets, and $\mathbb{A} \in \mathcal{A}$ be an action profile (i.e., a vector of all agents' actions). The reward for agent n playing arm $a_n^t \in A_n$ in round t is determined by function $u_n : \mathcal{A} \to [0, 1]$, which maps the actions of all agents to agent n's rewards $u_n(a_n^t; \mathbb{A}_{-n}^t)$.[2] Note that our algorithm and analyses also work for a time-varying reward function u_n^t. In addition, u_n^t can be determined in either an oblivious way or a non-oblivious (i.e., adaptive) way, corresponding to the oblivious adversary or the non-oblivious adversary in the single-agent bandits. In an oblivious way, $\{u_n^t\}_{t>0}$ is chosen at the beginning of the game, while in a non-oblivious way, each u_n^t is determined conditioned on all the agents' actions in the past.

One of the main differences between multi-agent bandits and single-agent bandits is the measurability of the rewards. If we are in the single-agent bandits, regardless of whether u_n^t is determined obliviously or non-obliviously, the reward of each arm in each round t is determined at the beginning of that round, before the agent plays an action. However, in the multi-agent bandits, as the reward of each arm for each agent is conditioned on other agents' actions, the reward of each arm in each round cannot be determined until all agents have played an action in that round.

Let $\mathcal{U} := \{u_1, \ldots, u_N\}$ be the set of reward functions for N agents. Note that neither $\mathcal{N}$ nor $\mathcal{U}$ is a prior knowledge to each agent, and each agent n only knows in advance her own set of arms A_n.

In each round $t = 1, \ldots, T$, each agent $n \in \mathcal{N}$ can use a *mixed* strategy to play an arm $a_n^t \in A_n$ according to a probability distribution over arms $P_n^t := \{p_a^t : \forall a \in A_n\}$, i.e., play arm $a \in A_n$ with probability p_a^t. Then, each agent n can only observe her own instantaneous reward $X_n^t := u_n(a_n^t; \mathbb{A}_{-n}^t)$.[3] Both the actions and the number of other agents cannot be observed. The objective of each agent is to accumulate as many rewards as possible over T rounds.

3.2 PROBLEM FORMULATION

As each agent has little knowledge about the environment, it is inevitable for each agent to suffer a *regret*, i.e., the loss of rewards for not playing the optimal arm in hindsight that returns the highest cumulative rewards. In bandit problems, the problem of maximizing the cumulative reward is always converted to the problem of minimizing the regret. The notion of regret has many forms. The most oft-used regret in the bandit literature is the *external regret* Cesa-Bianchi and Lugosi [2006]. Let $\mathbf{1}[a_n^t = a]$ be the indicator function that returns 1 if a is the played arm in round t and 0 otherwise. The external regret $R_n^{\text{ext}}(T)$ for agent n compares the cumulative reward of a learning algorithm with that of a set of competitors that always play a fixed arm up to round T, which is defined as follows:

$$R_n^{\text{ext}}(T) := \max_{a'\in A_n} \sum_{t=1}^{T} u_n(a'; \mathbb{A}_{-n}^t) - \sum_{t=1}^{T} \sum_{a\in A_n} \mathbf{1}[a_n^t = a] u_n(a; \mathbb{A}_{-n}^t),$$

However, only minimizing the external regret cannot guarantee the plays of agents will reach an equilibrium. Therefore, we need a strictly stronger notion of regret that is the *internal regret*, which compares the actions of an agent in a pair-wise manner:

$$R_n^{\text{int}}(T) := \max_{a,a'\in A_n} \sum_{t=1}^{T} r_{(a,a'),n}^t, \tag{1}$$

where

$$r_{(a,a'),n}^t := \mathbf{1}[a_n^t = a] \left(u_n(a'; \mathbb{A}_{-n}^t) - u_n(a; \mathbb{A}_{-n}^t) \right)$$

[2] $(a_n^t; \mathbb{A}_{-n}^t)$ is an abbreviation of $\mathbb{A}^t := (a_1^t, \ldots, a_n^t, \ldots, a_N^t)$ with a highlight of agent n's action a_n against other agents' actions.

[3] For the convenience of algorithm description and analysis, we sometimes use an equivalent notion called the instantaneous loss, i.e., $1 - X_n^t$, and denote by $y_{n,a}^t := 1 - u_n(a; \mathbb{A}_{-n}^t)$ the instantaneous loss function if agent n plays $a \in A_n$.

is the instantaneous regret for agent n of having played arm a instead of arm a' in round t. As proved in Hart and Mas-Colell [2000], Cesa-Bianchi and Lugosi [2006], by minimizing the internal regret for each agent, their empirical joint distributions of plays converge to an ϵ-correlated equilibrium, which is defined as follows.

Definition 3.1. Let $\mathbf{P}$ be a joint probability distribution over $\mathcal{A}$. We say $\mathbf{P}$ is an ϵ-correlated equilibrium if the expected incentive for each agent n to deviate from action a to any other action $a' \in A_n$ is no more than $\epsilon \geq 0$, i.e., $\forall n \in \mathcal{N}$, we have

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

Note that $\mathbf{P}$ is the joint distribution, not the product distribution, which is the difference between the correlated equilibrium and the Nash equilibrium. When $\epsilon = 0$, $\mathbf{P}$ is the correlated equilibrium, which is more general than the well-known Nash equilibrium, as the correlated equilibrium does not require independence among actions. To give an intuition about the ϵ-correlated equilibrium, consider a case of congestion control in computer networks where a *mediator* (e.g., a router or switch) draws an action profile from $\mathbf{P}$ and privately recommends each action (e.g., the packet sending rate) to the corresponding host. If no host has an incentive of more than ϵ to choose a different action, provided that other hosts follow the mediator's recommendation, then ϵ yields an ϵ-correlated equilibrium. Our objective is to achieve an ϵ-correlated equilibrium without a mediator by minimizing the internal regret for each agent.

In this paper, we consider a more general notion of regret, called the *swap regret* Blum and Mansour [2007], which can unify both the external regret and internal regret into the same framework by a swap function $F_n : A_n \rightarrow A_n$ that takes $a \in A_n$ as input and outputs $a' \in A_n$. Let $\mathcal{F}$ be a finite set of F_n. Then, the instantaneous swap regret for agent n with $\mathcal{F}$ up to round T is defined as follows:

$$R_n^{\text{swa}}(T,\mathcal{F}) = \max_{F\in\mathcal{F}} \sum_{t=1}^{T} \sum_{a\in A_n} \mathbf{1}[a_n^t = a]\left(u_n(F(a);\mathbb{A}_{-n}^t) - u_n(a;\mathbb{A}_{-n}^t)\right). \tag{3}$$

We can boil down the swap regret to the external regret by letting $\mathcal{F}$ be a set of K_n functions such that for any $a \in A_n$, $F_a \in \mathcal{F} : A_n \rightarrow a$. Similarly, the internal regret can be obtained by letting $\mathcal{F}$ be a set of $K_n(K_n - 1)$ functions such that for any pair of $a, a' \in A_n$, we have $F_{(a,a')}(a) = a'$ and $F_{(a,a')}(a'') = a''$ for any other $a'' \in A_n$. Thus, by minimizing the swap regret of a learning algorithm for a general $\mathcal{F}$ of any possible mappings F, we can show that the learning algorithm has a bounded performance gap from a broader range of competitors. We can also minimize the internal and external regrets at the same time, and achieve the ϵ-correlated equilibrium for all agents.

4 THE LCE-IX ALGORITHM

The LCE-IX algorithm adopts the swap-regret-minimizing framework introduced by Blum and Mansour [2007] and calls K_n Exp3-IX algorithms [Kocák et al., 2014, Neu, 2015] as subroutines. Each subroutine maintains a meta-distribution, and the action selection probability is calculated from the meta-distributions. The observed reward or loss will be assigned proportionally to each subroutine for updating the meta-distributions.

For each agent n, we define a meta-distribution $Q_a^t := \{q_{a,a'}^t : \forall a' \in A_n\}$ for each arm $a \in A_n$ such that $q_{a,a'}^t \in [0,1]$ and $\sum_{a'\in A_n} q_{a,a'}^t = 1$. Denote by $\mathbb{Q}_n^t := [Q_a^t]_{a\in A_n}$ the $K_n \times K_n$ matrix with each row being Q_a^t. Then, we determine the sample distribution P_n^t by solving the following equations:

$$P_n^t = P_n^t \mathbb{Q}_n^t, \tag{4}$$

where P_n^t is a row vector of $p_a^t, \forall a \in A_n$ and $\sum_{a\in A_n} p_a^t = 1$. That is, for each $a \in A_n$, we have $p_a^t = \sum_{a'\in A_n} p_{a'}^t q_{a',a}^t$, which is similar to the calculation of the stationary distribution of a Markov process with the transition matrix being $\mathbb{Q}_n^t$. The intuition behind (4) is to make the probability of playing arm $a' \in A_n$ directly according to P_n^t be equivalent to the probability of first sampling any arm $a \sim P_n^t$ and then playing a' according to Q_a^t.

The suffix 'IX' of LCE-IX stands for implicit exploration, which is justified by the γ_t-biased reward estimator defined as follows. Denote by $Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a']p_a^t q_{a,a'}^t}{p_{a'}^t}(1 - X_n^t)$ the loss of arm a' observed by subroutine a. Let γ_t be a non-negative and non-increasing parameter over time t. We then define the γ_t-biased estimated loss for $Y_{a,a'}^t$ as follows:

$$\hat{Y}_{a,a'}^t := \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma_t}.$$

This bias factor γ_t is introduced in Kocák et al. [2014], Neu [2015], which is used to smooth the meta-distributions so that the arms with low rewards in the past can still be chosen occasionally for exploration.

In addition, we also consider the situation where T may not be known a priori. Thus, we consider a non-increasing learning rate η_t, and the update rule for each meta-distribution is defined as follows:

$$q_{a,a'}^{t+1} = \frac{\exp\left(-\eta_{t+1}\hat{L}_{a,a'}^t\right)}{\sum_{a''\in A_n} \exp\left(-\eta_{t+1}\hat{L}_{a,a''}^t\right)}. \tag{5}$$

By the above modification, we have the LCE-IX algorithm described in Alg. 1.

Algorithm 1 The LCE-IX algorithm

```
Input: $n, A_n, \eta_t$
// Initialization
Set $q^1_{a,a'} = \frac{1}{K_n}$ and $\hat{L}^0_{a,a'} = 0, \forall a, a' \in A_n$
for $t = 1, \dots, T$ do
    // Compute the sample distribution, play arms and observe rewards
    Calculate $P^t_n$ based on (4)
    Play an arm $a^t_n \sim P^t_n$
    Observe reward $X^t_n$
    // Update each meta-distribution
    for $a \in A_a$ do
        $Y^t_{a,a'} := \frac{\mathbf{1}[a^t_n = a'] p^t_a q^t_{a,a'}}{p^t_{a'}}(1 - X^t_n), \forall a' \in A_n$
        $\hat{Y}^t_{a,a'} := \frac{Y^t_{a,a'}}{q^t_{a,a'} + \gamma_t}, \forall a' \in A_n$
        $\hat{L}^t_{a,a'} = \hat{L}^{t-1}_{a,a'} + \hat{Y}^t_{a,a'}, \forall a' \in A_n$
        Calculate $Q^{t+1}_a$ based on (5)
    end for
end for
```

5 ANALYTICAL RESULTS FOR LCE-IX

5.1 REGRET BOUND

As the regret analysis is for each individual agent $n \in \mathcal{N}$, without confusion, we drop the subscript n in some notations for brevity. Let $\mathcal{G}_t$ denote the σ-algebra generated by the history information of all agents up to round t, i.e., $\mathcal{G}_t := \sigma\left(\{a^1_n, r^1_n, \dots, a^t_n, r^t_n\}_{n \in \mathcal{N}}\right)$. Denoted by $\tilde{Y}_{a,a'} := \mathbf{1}[a^t_n = a] y^t_{a'}$ the swapped loss from a to a', where $y^t_{a'} := 1 - X^t_n$. We first state a novel concentration bound for the γ_t-biased loss estimator used in LCE-IX, which shows that the cumulative gap between the biased loss estimator $\hat{Y}^t_{a,a'}$ and the swapped loss $\tilde{Y}^t_{a,a'}$ for each agent $n \in \mathcal{N}$ is bounded with a high probability.

Lemma 5.1. *Let $\delta \in (0,1)$ and let $\beta^t_{a,a'}$ be nonnegative and non-increasing (over time t) $\mathcal{G}_{t-1}$-measurable random variables (i.e., given $\mathcal{G}_{t-1}$, $\beta^t_{a,a'}$ is determined) satisfying $\beta^t_{a,a'} \le 2\gamma_t$ for all pairs $a, a' \in A_n$. With probability at least $1 - \delta$, we have the following inequality held:*

$$\sum_{t=1}^{T} \sum_{a \in A_n} \sum_{a' \in A_n} \beta^t_{a,a'} \left(\hat{Y}^t_{a,a'} - \tilde{Y}^t_{a,a'}\right) \le 2\log(\frac{2}{\delta}) \,. \quad (6)$$

Proof Sketch. We only give a proof sketch here, and the detailed proof can be found in Appendix A in the supplementary materials. First, we construct a sequence of random variables $\{Z_t\}_{t \ge 0}$, where $Z_t := \exp\left\{\sum_{s=1}^{t} \beta^s_{a,a'} \sum_{a \in A_n} \sum_{a' \in A_n} \left(\hat{Y}^s_{a,a'} - \tilde{Y}^s_{a,a'}\right)\right\}$ for $t > 0$ and $Z_0 = 1$, and then prove that $\{Z_t\}_{t \ge 0}$ is a supermartingale with respect to filtration $\{\mathcal{G}_t\}_{t \ge 0}$, i.e., $\mathbf{E}\left[Z_t | \mathcal{G}_{t-1}\right] \le Z_{t-1}$. Finally, the lemma follows the Markov inequality. □

The proof for Lemma 5.1 is refined beyond the IX concentration bounds studied in Neu [2015]. The original approach used in Neu [2015] is for external regret, but swap regret is only equivalent to the sum of the external regret for subroutine algorithms in expectation. Therefore, we cannot simply adapt their concentration bound to analyze the instantaneous swap regret. In addition, in the original concentration bound, the probability is taken with respect to only one agent's randomness, which is not suitable for the MAB-UG setting, as the reward/loss for each agent is dependent on all agents' actions. To address the issue, Lemma 5.1 considers the K_n subroutines as a whole, and proves a supermartingale between the sum of IX loss estimators for each meta-distribution and the general swapped loss with respect to all agents' randomness. The following Lemma is a direct result of Lemma 5.1, which is essential for the swap-regret analysis.

Lemma 5.2. *Let $\delta \in (0,1)$. With probability at least $1 - \delta$, the following inequalities hold simultaneously:*

$$\sum_{t=1}^{T} \sum_{a \in A_n} \sum_{a' \in A_n} \eta_t \left(\hat{Y}^t_{a,a'} - \tilde{Y}^t_{a,a'}\right) \le 2\log(\frac{2}{\delta}), \quad (7)$$

and for any $F \in \mathcal{F}$,

$$\sum_{t=1}^{T} \sum_{a \in A_n} \left(\hat{Y}^t_{a,F(a)} - \tilde{Y}^t_{a,F(a)}\right) \le \frac{1}{\gamma_T} \log(\frac{2 \cdot K_n^{K_n}}{\delta}). \quad (8)$$

Proof. (7) is obtained by invoking Lemma 5.1 with $\beta^t_{a,a'} := \eta_t$ for all $a, a' \in A_n$. (8) is obtained by invoking Lemma 5.1 with $\beta^t_{a,a'} := 2\gamma_T \mathbf{1}[a' = F(a)]$ for all $a, a', F(a) \in A_n$ and applying the union bound over all $F \in \mathcal{F}$ for at most $|\mathcal{F}| = K_n^{K_n}$ swap functions. □

The regret defined in (3) for each agent $n \in \mathcal{N}$ playing the LCE-IX algorithm is guaranteed by the following theorem.

Theorem 5.3. *Let $\delta \in (0,1)$. With probability at least $1 - \delta$, $\eta_t = \sqrt{\frac{\log(K_n)}{t}}$ and $\gamma_t = \eta_t/2$, the instantaneous swap regret for playing the LCE-IX algorithm over T rounds is bounded as follows*

$$R^{\text{swap}}_n(T, \mathcal{F}) \le 4K_n\sqrt{T\log(K_n)} + \left(2 + K_n\sqrt{\frac{T}{\log(K_n)}}\right)\log(\tfrac{2}{\delta}). \quad (9)$$

When $\eta_t = \sqrt{\frac{\log(K_n) + \log(K_n/\delta)}{t}}$ and $\gamma_t = \eta_t/2$, the above bound becomes

$$R^{\text{swap}}_n(T, \mathcal{F}) \le 3K_n\sqrt{T(\log(K_n) + \log(2K_n/\delta))} + \log(\tfrac{2}{\delta}). \quad (10)$$

Proof Sketch. The regret defined in (3) can be rewritten in the loss form and can be decomposed as follows:

$$R_n^{\text{swap}}(T,\mathcal{F}) \leq \underbrace{\sum_{a\in A_n}(L_a^T - \sum_{a\in A_n}\hat{L}_a^T)}_{=:(\mathrm{a})} + \underbrace{\sum_{a\in A_n}(\hat{L}_a^T - \hat{L}_{a,F(a)}^T)}_{=:(\mathrm{b})} + \underbrace{\sum_{a\in A_n}(\hat{L}_{a,F(a)}^T - \tilde{L}_{a,F(a)}^T)}_{=:(c)},$$

where $L_a^T := \sum_{t=1}^{T}\sum_{a'\in A_n} Y_{a,a'}^t$ and $\hat{L}_a^t := \sum_{t=1}^{T}\sum_{a'\in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t$ are the cumulative instantaneous and estimated loss allocated to meta-distribution Q_a^t over T rounds, respectively.

Then, we can bound (a) by $\sum_{a\in A_n}\sum_{t=1}^{T}\gamma_t \sum_{a'\in A_n}\hat{Y}_{a,a'}^t$, which is a straightforward result by the definition of $\hat{Y}_{a,a'}^t$ and (b) is bounded by $\frac{K_n\log(K_n)}{\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{a\in A_n}\sum_{a'\in A_n}\hat{Y}_{a,a'}^t$ by a refined analysis for the exponential-weighting technique. Finally, invoking Lemma 5.2 can bound (c) and term $\sum_{t=1}^{T}(\gamma_t + \frac{\eta_t}{2})\sum_{a\in A_n}\sum_{a'\in A_n}\hat{Y}_{a,a'}^t$. The detailed proof can be found in Appendix B in the supplementary materials.

□

Note that it is not required for all agents to play LCE-IX at the same time to guarantee Theorem 5.3. The value of η_t for the bound in (9) is independent of δ, which means the bound holds for all δ. On the other hand, the high-probability bound in (10) is improved when the algorithm can use a fixed confidence level δ to tune its parameters. The former bound, however, is useful for deriving an expected swap-regret bound as shown in the following corollary.

Corollary 5.4. *With $\eta_t = \sqrt{\frac{\log(K_n)}{t}}$ and $\gamma_t = \eta_t/2$, the expected swap regret is bounded as follows:*

$$\mathbf{E}[R_n^{\text{swap}}(T,\mathcal{F})] \leq 4K_n\sqrt{T\log(K_n)} + 2K_n\sqrt{\frac{T}{\log(K_n)}} + 4$$

Proof. Let $W := \frac{R_n^T(T,\mathcal{F}) - 4K_n\sqrt{T\log(K_n)}}{2 + K_n\sqrt{\frac{T}{\log(K_n)}}}$. By (9), we have that $\Pr(W > \log(\frac{2}{\delta})) \leq \delta$. Then, integrating the tail gives $\mathbf{E}[W] \leq \int_0^1 \frac{2}{\delta}\Pr(W > \log(\frac{2}{\delta}))d\delta \leq 2$. □

The expected swap regret is upper bounded in $O(K_n\sqrt{T\log(K_n)})$, which we claim is near-optimal because there is a gap of $O(\sqrt{K_n})$ from the lower bound of $\Omega(\sqrt{K_n T\log(K_n)})$ proved in Ito [2020]. However, the lower bound there is tight for the full-information model, but may not be tight with the bandit feedback.

5.2 CONVERGENCE TO CORRELATED EQUILIBRIA

If every agent involved in the game plays the LCE-IX algorithm at the same time, the following theorem guarantees that the empirical distribution $\hat{\mathbf{P}}^T$ of the joint actions converges to an ϵ-correlated equilibrium.

Theorem 5.5. *If every agent $n \in \mathcal{N}$ plays the LCE-IX algorithm for T rounds, then the empirical distribution of the joint actions played by all agents $\hat{\mathbf{P}}^T$ is an ϵ-correlated equilibrium with probability at least $1-\delta$, where $\epsilon = O(\max_{n\in\mathcal{N}} K_n\sqrt{\frac{\log(K_n N/\delta)}{T}})$. When $T \to \infty$, the empirical distribution of the joint actions converges to the set of correlated equilibria almost surely.*

Proof. Let $\delta' > 0$. By (10), with probability $1-\delta'$, $R_n^{\text{int}}(T) \leq 3K_n\sqrt{T(\log(K_n) + \log(2K_n/\delta'))} + \log\frac{1}{\delta'}$ for agent n. By using the union bound over all the N agents and letting $\delta' = \delta/N$, we have that with probability at least $1-\delta$: $\sum_{\mathbb{A}:a_n=a}\hat{\mathbf{P}}^T(\mathbb{K})\left(r_n\left(a';\mathbb{A}_{-n}\right) - r_n\left(\mathbb{A}\right)\right) = \frac{1}{T}R_n^{\text{int}}(T) \leq O(\max_{n\in\mathcal{N}} K_n\sqrt{\frac{\log(K_n N/\delta)}{T}})$. When $T \to \infty$, by the Borel-Cantelli Lemma, we have $\limsup_{T\to\infty}\frac{1}{T}R_n^{\text{int}}(T) \leq 0$ almost surely, which indicates the empirical distribution of joint actions converges to the set of correlated equilibria.

□

Solving the equation $\epsilon = O(\max_{n\in\mathcal{N}} K_n\sqrt{\frac{\log(K_n N/\delta)}{T}})$ for T implies that the empirical joint distribution $\hat{\mathbf{P}}^T$ for all agents meets the definition of an ϵ-correlated equilibrium for the unknown games after $T = \Omega(\max_{n\in\mathcal{N}}\frac{K_n^2\log(K_n N/\delta)}{\epsilon^2})$ rounds, i.e., the equilibrium is achieved.

5.3 TIME AND SPACE COMPLEXITY

In each round, each agent needs first to calculate P_n^t based on (4), which can be regarded as the calculation of a stationary distribution for the Markov process defined by Q_n^t, and can be achieved within $O(K_n^2)$ for K_n states [Feinberg and Chiu, 1987]. Then, each meta-distribution needs $O(K_n)$ time to be updated for K_n arms. Therefore, the time complexity for LCE-IX is $O(K_n^2)$. Regarding the space complexity, we need to maintain K_n meta-distributions for the LCE algorithm, and each meta-distribution requires $O(K_n)$ space for K_n arms, so the space complexity is $O(K_n^2)$.

6 NUMERICAL EXPERIMENTS

In this section, we compare LCE-IX with LCE (i.e., $\gamma_t = 0$) to show the effectiveness of the IX technique. We also compare with a recent algorithm with the full-information feedback called BM-Opt-Hedge [Chen and Peng, 2020]. The results are the average of 100 independent experiments.

We study a wireless medium access game between two wireless devices (i.e., two agents), where the two wireless devices are *hidden* from each other (i.e., each device cannot observe each other), and trying to access one unknown channel in each round. Each device has two options in each round, wait for the next round (W) or access in the current round (A). If a device chooses action W, it will receive a reward of 0.

If a device chooses to access (A), the device has an energy cost of 0.2. When only one device successfully accesses the medium, then this device will receive a reward of 0.8. If both devices choose action A, then there is a collision and the rewards for both devices are -0.2 due to the wasted transmission energy. The reward matrix (unknown to the agents) is shown in Table 2.

We assume that all the devices do not adopt the RTS/CTS mechanism, an oft-used technique to solve the hidden terminal problem, as it will introduce new problems among its control messages Sobrinho et al. [2005] and the game model still applies to the RTS/CTS message itself. Thus, it is quite challenging to improve the received rewards (i.e., the successful access to the channel) for both devices in a distributed way, as *both wireless devices are hidden terminals to each other so that they cannot observe the actions and rewards of each other and they do not know the total number of devices.*

Table 2: The reward matrix for the medium access game

	W	A
W	$(0, 0)$	$(0, 0.8)$
A	$(0.8, 0)$	$(-0.2, -0.2)$

As the swap regret is a generic performance measure, different swap functions can lead to different regret definitions. In this experiment, we show two metrics that reflect two different regret definitions. The first metric is the time-averaged reward, the gap of which from the optimal actions in hindsight reflects the external regret of an online learning algorithm. The other metric is the convergence to the ϵ-correlated equilibrium. Whether or not an ϵ-correlated equilibrium can be reached reflects whether an online learning algorithm can minimize its internal regret.

6.1 TIME-AVERAGED REWARD

To save space, we only show the time-averaged reward for all the considered algorithms, as it contains the equivalent information about the cumulative regrets or rewards. For example, if an algorithm has higher time-averaged rewards (or closer to the maximum rewards), then it also has higher cumulative rewards (or lower cumulative regrets).

To compare with a benchmark, we consider an adaptive access technique (denoted by `Ada`) with the prior knowledge about the number of all the hidden devices, which randomly accesses a channel with an initial probability $\frac{1}{2}$ for two devices. If a device fails, the access probability of that device is reduced by half; otherwise, the device uses the initial probability in the next round. `Ada` in our experiments can achieve better performance than the distributed coordination function of IEEE 802.11 used by current WiFi devices in real-world scenarios, as `Ada` in our experiments can set an appropriate initial probability to achieve high throughput. Therefore, `Ada` can be a good benchmark with partial prior knowledge to show the effectiveness of the swap-regret-minimizing algorithms.

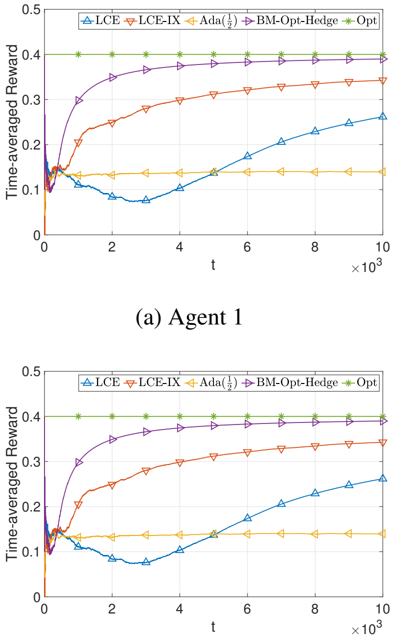
In addition, the maximum time-averaged reward (denoted by `Opt`) of 0.4 can be achieved, by a mediator (e.g., wireless access point) with full prior knowledge which either asks Agent 1 to play W and Agent 2 to play A, or asks Agent 1 to play A and Agent 2 to play W in each round. We show that LCE-IX can approach `Opt` quickly in a distributed fashion over time.

The time-averaged rewards of both agents in 1×10^4 rounds are shown in Fig. 2. As we can see, LCE-IX outperforms both LCE and `Ada`($\frac{1}{2}$) in terms of the faster convergence to `Opt`. This shows the effectiveness of the γ_t-biased estimator in smoothing the reward estimation so that the low-reward arm can still be explored occasionally. We can also see that BM-Opt-Hedge achieves the fastest result, but we note that BM-Opt-Hedge is with the full-information feedback.

6.2 CONVERGENCE TO THE ϵ-CORRELATED EQUILIBRIUM

The convergence of empirical distribution of joint actions played by the two agents in T rounds is shown in Fig. 3b, where (W,W) means both agents play action W, (W,A) means Agent 1 plays W and Agent 2 plays A and so on. We take the result of LCE-IX to explain the convergence to the correlated equilibrium. The final results in Fig. 3a are $\hat{\mathbf{P}}^T(W, W) = 0.0088$, $\hat{\mathbf{P}}^T(W, A) = 0.4501$, $\hat{\mathbf{P}}^T(A, W) = 0.4501$, and $\hat{\mathbf{P}}^T(A, A) = 0.091$. We can do a simple calculation to verify this empirical distribution is a correlated equilibrium ($\epsilon = 0$). For example, the expected incentives for Agent 1 to switch from W to A are $\hat{\mathbf{P}}^T(W, W) \cdot u_1(A, W) + \hat{\mathbf{P}}^T(W, A) \cdot u_1(A, A) - (\hat{\mathbf{P}}(W, W) \cdot u_1(W, W) + \hat{\mathbf{P}}^T(W, A) \cdot u_1(W, A)) = -0.08298 < 0$, showing that Agent 1 does not have incentives to switch from W to A when both agents follow the joint distribution $\hat{\mathbf{P}}^T$. In the same way, we can verify the empirical joint distribution $\hat{\mathbf{P}}^T$ is a correlated equilibrium for both agents.

Fig. 3b shows that LCE-IX has a faster convergence than LCE to a correlated equilibrium, as the empirical proba-



(a) Agent 1

(b) Agent 2

Figure 2: The time-averaged reward for both agents.

bilities of the optimal action pairs of (A, W) and (W, A) increase faster than that of LCE. This again shows the effectiveness of the γ-biased estimator in controlling the variation of the reward estimation.

On the other hand, with full-information feedback, BM-Opt-Hedge can achieve a faster convergence rate than LCE and LCE-IX. It will be our future interest to study whether the techniques of BM-Opt-Hedge can be applied to the bandit-feedback model to speed up the convergence rate.

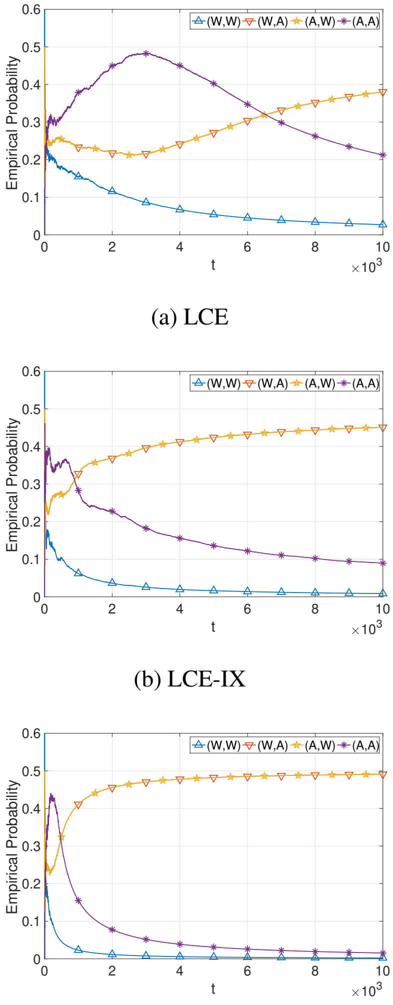
7 CONCLUSION

In this paper, with regard to the randomness of all agents' actions, we provided a high-probability bound for the instantaneous swap regret, which can further bound the expected swap regret. Furthermore, we conducted numerical experiments to verify the performance of LCE-IX.

Regarding future work, we will study the swap regret bounds for mirror descent algorithms, and aim to close the gap between the upper bound and the lower bound for swap regret.

Acknowledgements

We want to thank Prof. Nishant Mehta for his careful reading of the paper and pointing out an issue in Lemma 5.1, which has now been fixed.



(a) LCE

(b) LCE-IX

(c) BM-Opt-Hedge

Figure 3: The empirical distribution of joint actions by two agents in T rounds.

References

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal No-regret Learning for Correlated Equilibria in Multi-player General-sum Games. In *Proc. ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 736–749, 2022.

Jean-Yves Audibert and Sébastien Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *The Journal of Machine Learning Research (JMLR)*, 11:2785–2836, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Robert J Aumann. Subjectivity and Correlation in Randomized Strategies. *Journal of Mathematical Economics*, 1 (1):67–96, 1974.

Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Michał Misiurewicz, and Georgios Piliouras. Follow-the-Regularized-Leader Routes to Chaos in Routing Games. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, pages 925–935. PMLR, 18–24 Jul 2021.

Avrim Blum and Yishay Mansour. From External to Internal Regret. *Journal of Machine Learning Research (JMLR)*, 8(6), 2007.

George W Brown. Some Notes on Computation of Games Solutions. Technical report, RAND Corp Santa Monica CA, 1949.

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-stochastic Multi-player Multi-armed Bandits: Optimal Rate with Collision Information, Sublinear Without. In *Proc. Conference on Learning Theory (COLT)*, pages 961–987. PMLR, 2020.

Swapna Buccapatnam, Jian Tan, and Li Zhang. Information Sharing in Distributed Stochastic Bandits. In *Proc. IEEE Conference on Computer Communications (INFOCOM)*, pages 2605–2613. IEEE, 2015.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus Decentralized Exploration in Multi-agent Multi-armed Bandits. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 164–170, 2017.

Xi Chen and Binghui Peng. Hedging in Games: Faster Convergence of External and Swap Regrets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with Bandit Feedback in Potential Games. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 6372–6381, 2017.

Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated Dynamics and Robust Learning Procedures in Games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal No-regret Learning in General Games. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Abhimanyu Dubey et al. Cooperative Multi-agent Bandits with Heavy Tails. In *Proc. International Conference on Machine Learning (ICML)*, pages 2730–2739. PMLR, 2020.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-Optimal No-Regret Learning Dynamics for General Convex Games. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Brion N Feinberg and Samuel S Chiu. A Method to Calculate Steady-state Distributions of Large Markov Chains by Aggregating States. *Operations Research*, 35(2):282–290, 1987.

Sergiu Hart and Andreu Mas-Colell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed Exploration in Multi-armed Bandits. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 854–862, 2013.

Shinji Ito. A Tight Lower Bound and Efficient Reduction for Swap Regret. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18550–18559, 2020.

Kevin Jamieson and Robert Nowak. Best-arm Identification Algorithms for Multi-armed Bandits in the Fixed Confidence Setting. In *Proc. Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-Learning–A Simple, Efficient, Decentralized Algorithm for Multiagent RL. In *Proc. ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient Learning by Implicit Exploration in Bandit Problems with Side Observations. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 613–621, 2014.

Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative Learning of Stochastic Bandits over a Social Network. *IEEE/ACM Transactions on Networking (TON)*, 26(4):1782–1795, 2018.

Walid Krichene, Benjamin Drighès, and Alexandre M Bayen. Online Learning of Nash Equilibria in Congestion Games. *SIAM Journal on Control and Optimization*, 53 (2):1056–1081, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Michael L Littman. Markov Games as A Framework for Multi-agent Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 157–163, 1994.

Keqin Liu and Qing Zhao. Distributed Learning in Multi-armed Bandit with Multiple Players. *IEEE Transactions on Signal Processing (TSP)*, 58(11):5667–5681, 2010.

David H Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. Learning in Nonzero-Sum Stochastic Games with Potentials. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, pages 7688–7699. PMLR, 18–24 Jul 2021.

Ming Min and Ruimeng Hu. Signatured Deep Fictitious Play for Mean Field Games with Common Noise. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, pages 7736–7747. PMLR, 18–24 Jul 2021.

Heinrich H Nax, Maxwell N Burton-Chellew, Stuart A West, and H Peyton Young. Learning in A Black Box. *Journal of Economic Behavior & Organization*, 127:1–15, 2016.

Gergely Neu. Explore No More: Improved High-probability Regret Bounds for Non-Stochastic Bandits. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios 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)*, volume 30, pages 5872–5882, 2017.

Julia Robinson. An Iterative Method of Solving a Game. *Annals of Mathematics*, pages 296–301, 1951.

João L Sobrinho, Roland De Haan, and José M Brazio. Why RTS-CTS is Not Your Ideal Wireless LAN Multiple Access Protocol. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, volume 1, pages 81–87. IEEE, 2005.

Gilles Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris Sud-Paris XI, 2005.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based Distributed Stochastic Bandit Algorithms. In *Proc. International Conference on Machine Learning (ICML)*, pages 19–27. PMLR, 2013.

Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust Multi-agent Multi-armed Bandits. In *Proc. International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, pages 161–170, 2021.

Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global Convergence of Policy Gradient for Linear-Quadratic Mean-Field Control/Game in Continuous Time. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, pages 10772–10782. PMLR, 18–24 Jul 2021.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning While Playing in Mean-Field Games: Convergence and Optimality. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, pages 11436–11447. PMLR, 18–24 Jul 2021.

Julian Zimmert and Yevgeny Seldin. An Optimal Algorithm for Stochastic and Adversarial Bandits. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475. PMLR, 2019.

We start by introducing the notations that will be used in the proofs of Lemma 5.1 and Theorem 5.3. As the proofs are for each individual agent n, without confusion, we drop the subscript n in some notations for brevity.

Recall that $\mathcal{G}_t$ the σ-algebra generated by the history information of all agents till round t, i.e., $\mathcal{G}_t := \sigma\left(\{a_n^1, r_n^1, \ldots, a_n^t, r_n^t\}_{n\in\mathcal{N}}\right)$ and let $\mathbf{E}_t[\cdot] := \mathbf{E}[\cdot|\mathcal{G}_t]$ be the expectation conditioned on the history information by the end of round t. Recall that $y_a^t := 1 - u_n^t(a; \mathbb{A}_{-n}^t)$ is the instantaneous loss function if agent n plays arm $a \in A_n$ in round t, and thus $Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a'] p_a^t q_{a,a'}^t y_{a'}^t}{p_{a'}^t}$ and $\hat{Y}_{a,a'}^t = \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma_t}$. Denote by $\hat{L}_a^t := \sum\limits_{t=1}^{T} \sum\limits_{a'\in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t$ and $L_a^T := \sum\limits_{t=1}^{T} \sum\limits_{a'\in A_n} Y_{a,a'}^t$.

A PROOF OF LEMMA 5.1

Proof. We can decompose $\sum\limits_{t=1}^{T} \sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t \left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right)$ as follows:

$$\sum_{t=1}^{T} \sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right) = \sum_{t=1}^{T} \sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \left(\hat{Y}_{a,a'}^t - p_a^t y_{a'}^t\right) + \sum_{t=1}^{T} \sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \left(p_a^t y_{a'}^t - \tilde{Y}_{a,a'}^t\right). \tag{11}$$

We first bound the first term in (11) by proving that the process $\{Z_t\}_{t\geq 0}$, where $Z_t := \exp\left\{\sum\limits_{s=1}^{t} \sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^s \left(\hat{Y}_{a,a'}^s - p_a^s y_{a'}^s\right)\right\}$ for $t > 0$ and $Z_0 = 1$, is a supermartingale with respect to filtration $\{\mathcal{G}_t\}_{t\geq 0}$ for all $a \in A_n$, i.e., $\mathbf{E}\left[Z_t|\mathcal{G}_{t-1}\right] \leq Z_{t-1}$. Denote by $\mathbb{A}_{-n}^t$ the actions of all agents except agent n in round t. Then, we have that

$$\begin{aligned}
&\mathbf{E}_{t-1}\left[\exp\left\{\sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \left(\hat{Y}_{a,a'}^t - p_a^t y_{a'}^t\right)\right\}\right] = \mathbf{E}_{t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t \hat{Y}_{a,a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t p_a^t y_{a'}^t\right\}}\right] \\
&= \mathbf{E}_{t-1}\left[\mathbf{E}_{t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t \hat{Y}_{a,a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t p_a^t y_{a'}^t\right\}} \mid \mathbb{A}_{-n}^t\right]\right] = \mathbf{E}_{t-1}\left[\frac{\mathbf{E}_{t-1}\left[\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t \hat{Y}_{a,a'}^t\right\} \mid \mathbb{A}_{-n}^t\right]}{\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t p_a^t y_{a'}^t\right\}}\right],
\end{aligned} \tag{12}$$

where the third equality is due to the law of total expectation, and the last inequality is due to that $y_{a'}^t$ is determined given $\mathbb{A}_{-n}^t$ and $\beta_{a,a'}^t$ is $\mathcal{G}_{t-1}$-measurable.

Denote by $\mathbf{E}_{n,t-1}[\cdot] := \mathbf{E}_{t-1}\left[\cdot \mid \mathbb{A}_{-n}^t\right]$. Then, we show that $\mathbf{E}_{n,t-1}\left[\exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t \hat{Y}_{a,a'}^t\right\}\right] \leq \exp\left\{\sum\limits_{a\in A_n} \sum\limits_{a'\in A_n} \beta_{a,a'}^t p_a^t y_{a'}^t\right\}$ as follows:

$$\begin{aligned}
&\mathbf{E}_{n,t-1}\left[\exp\left\{\sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \hat{Y}_{a,a'}^t\right\}\right] = \mathbf{E}_{n,t-1}\left[\exp\left\{\sum_{a\in A_n} \sum_{a'\in A_n} \beta_{a,a'}^t \frac{p_a^t \mathbf{1}[a_n^t = a'] q_{a,a'}^t y_{a'}^t}{p_{a'}^t (q_{a,a'}^t + \gamma_t)}\right\}\right] \\
&\leq \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} \exp\left\{\beta_{a,a'}^t \frac{\mathbf{1}[a_n^t = a'] y_{a'}^t}{q_{a,a'}^t + \gamma_t}\right\}\right] \leq \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} \exp\left\{\frac{\beta_{a,a'}^t}{2\gamma_t} \frac{2\gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t}{q_{a,a'}^t + \gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t}\right\}\right] \\
&= \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} \exp\left\{\frac{\beta_{a,a'}^t}{2\gamma_t} \frac{2\gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t}{q_{a,a'}^t + \gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t}\right\}\right] = \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} \exp\left\{\frac{\beta_{a,a'}^t}{2\gamma_t} \frac{2\gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t / q_{a,a'}^t}{1 + \gamma_t \mathbf{1}[a_n^t = a'] y_{a'}^t / q_{a,a'}^t}\right\}\right] \\
&\leq \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} \exp\left\{\log(1 + \beta_{a,a'}^t \mathbf{1}[a_n^t = a'] y_{a'}^t / q_{a,a'}^t)\right\}\right] = \mathbf{E}_{n,t-1}\left[\sum_{a\in A_n} \sum_{a'\in A_n} \frac{p_a^t q_{a,a'}^t}{p_{a'}^t} (1 + \beta_{a,a'}^t \mathbf{1}[a_n^t = a'] y_{a'}^t / q_{a,a'}^t)\right].
\end{aligned}$$

where the first inequality is due to Jensen's inequality, the second inequality is due to that $0 \leq \mathbf{1}[a_n^t = a']q_{a,a'}^t y_{a'}^t \leq 1$, the third inequality is due to the fact that $\frac{z}{1+z/2} \leq \log(1+z)$ for all $z > 0$, and the last inequality is due to the inequality $x\log(1+y) \leq \log(1+xy)$ for all $y > -1$ and $x \in [0,1]$. The last term in above equation can be further processed as follows:

$$\mathbf{E}_{n,t-1}\left[\sum_{a\in A_n}\sum_{a'\in A_n}\frac{p_a^t q_{a,a'}^t}{p_{a'}^t}(1+\beta_{a,a'}^t\mathbf{1}[a_n^t = a']y_{a'}^t/q_{a,a'}^t)\right] = \mathbf{E}_{n,t-1}\left[1+\sum_{a\in A_n}\sum_{a'\in A_n}\frac{p_a^t q_{a,a'}^t}{p_{a'}^t}\beta_{a,a'}^t\mathbf{1}[a_n^t = a']y_{a'}^t/q_{a,a'}^t)\right]$$

$$= \mathbf{E}_{n,t-1}\left[1+\sum_{a\in A_n}\sum_{a'\in A_n}\frac{p_a^t}{p_{a'}^t}\beta_{a,a'}^t\mathbf{1}[a_n^t = a']y_{a'}^t)\right] = 1+\sum_{a\in A_n}\sum_{a'\in A_n}p_a^t\beta_{a,a'}^t y_{a'}^t \leq \exp\left\{\sum_{a\in A_n}\sum_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\},$$

where the inequality is due to $1+x \leq \exp\{x\}$ for any $x \in \mathbb{R}$. Therefore, we have shown that $\mathbf{E}_{n,t-1}\left[\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\hat{Y}_{a,a'}^t\right\}\right] \leq \exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}$, which indicates that (12) is bounded by 1. Thus,

$$\mathbf{E}_{t-1}\left[Z_t\right] = \mathbf{E}_{t-1}\left[\exp\left\{\sum_{a\in A_n}\sum_{a'\in A_n}\beta_{a,a'}^t\left(\hat{Y}_{a,a'}^t - p_a^t y_{a'}^t\right)\right\}\right]\cdot Z_{t-1} \leq Z_{t-1},$$

which shows that $\{Z_t\}_{t\geq 0}$ is a supermartingale with respect to filtration $\{\mathcal{G}_t\}_{t\geq 0}$. Thus, we have $\mathbf{E}\left[Z_T\right] \leq \mathbf{E}\left[Z_{T-1}\right] \leq \ldots \leq \mathbf{E}\left[Z_0\right] = 1$. By the Markov inequality, we have

$$\Pr\left(\sum_{t=1}^T\beta_{a,a'}^t\sum_{a\in A_n}\sum_{a'\in A_n}\left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right)\geq\epsilon\right) \leq \mathbf{E}\left[\exp\left\{\sum_{t=1}^T\beta_{a,a'}^t\sum_{a\in A_n}\sum_{a'\in A_n}\left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right)\right\}\right]\cdot\exp\{-\epsilon\}$$
$$\leq \exp\{-\epsilon\}.$$

Next, we bound the second item in (11) in a similar way by proving $\{S_t\}_{t\geq 0}$ is a supermartingale sequence, where $S_t := \exp\left\{\sum\limits_{s=1}^t\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^s\left(p_a^s y_{a'}^s - \tilde{Y}_{a,a'}^s\right)\right\}$ and $S_0 = 1$. Recall that $\tilde{Y}_{a,a'}^t := \mathbf{1}[a_n^t = a]y_{a'}^t$. Thus, we have that

$$\mathbf{E}_{t-1}\left[\exp\left\{\sum_{a\in A_n}\sum_{a'\in A_n}\beta_{a,a'}^t\left(p_a^t y_{a'}^t - \tilde{Y}_{a,a'}^t\right)\right\}\right] = \mathbf{E}_{t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\tilde{Y}_{a,a'}^t\right\}}\right] = \mathbf{E}_{t-1}\left[\mathbf{E}_{n,t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\tilde{Y}_{a,a'}^t\right\}}\right]\right]$$

$$= \mathbf{E}_{t-1}\left[\mathbf{E}_{n,t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\mathbf{1}[a_n^t = a]y_{a,a'}^t\right\}}\right]\right] = \mathbf{E}_{t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\mathbf{E}_{n,t-1}\left[\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\mathbf{1}[a_n^t = a]y_{a,a'}^t\right\}\right]}\right]$$

$$\leq \mathbf{E}_{t-1}\left[\mathbf{E}_{n,t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t\mathbf{1}[a_n^t = a]y_{a,a'}^t\right\}}\right]\right] = \mathbf{E}_{t-1}\left[\frac{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a'}^t\right\}}{\exp\left\{\sum\limits_{a\in A_n}\sum\limits_{a'\in A_n}\beta_{a,a'}^t p_a^t y_{a,a'}^t\right\}}\right] = 1,$$

where the inequality is due to Jensen's inequality. Thus, we have $\mathbf{E}[S_T] \leq \mathbf{E}[S_{T-1}] \leq \ldots \leq \mathbf{E}[S_0] = 1$. By the Markov inequality, we have that

$$\Pr\left(\sum_{t=1}^T\beta_{a,a'}^t\sum_{a\in A_n}\sum_{a'\in A_n}\left(p_a^t y_{a'}^t - \tilde{Y}_{a,a'}^t\right)\geq\epsilon\right)\leq\exp\{-\epsilon\}.$$

Then, by the union bound, we have that

$$\Pr\left(\sum_{t=1}^T\sum_{a\in A_n}\sum_{a'\in A_n}\beta_{a,a'}^t\left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right)\geq 2\epsilon\right)\leq\Pr\left(\sum_{t=1}^T\beta_{a,a'}^t\sum_{a\in A_n}\sum_{a'\in A_n}\left(\hat{Y}_{a,a'}^t - \tilde{Y}_{a,a'}^t\right)\geq\epsilon\right)+\Pr\left(\sum_{t=1}^T\beta_{a,a'}^t\sum_{a\in A_n}\sum_{a'\in A_n}\left(p_a^t y_{a'}^t - \tilde{Y}_{a,a'}^t\right)\geq\epsilon\right)\leq 2\exp\{-\epsilon\},$$

and the lemma follows by solving $2\exp\{-\epsilon\} = \delta$ for ϵ.

□

B PROOF OF THEOREM 5.3

Proof. By the relationship between P_n^t and Q_a^t, we have the following equation held:

$$\begin{aligned}\sum_{a\in A_n} L_a^T &= \sum_{a\in A_n}\sum_{t=1}^{T}\sum_{a'\in A_n} Y_{a,a'}^t = \sum_{t=1}^{T}\sum_{a'\in A_n}\sum_{a\in A_n}\frac{\mathbf{1}[a_n^t=a']p_a^t q_{a,a'}^t}{p_{a'}^t} y_{a'}^t \\ &= \sum_{t=1}^{T}\sum_{a'\in A_n}\mathbf{1}[a_n^t=a']y_{a'}^t = \sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t=a]y_a^t,\end{aligned} \tag{13}$$

The regret defined in (3) can be rewritten in the loss form and can be decomposed as follows:

$$\begin{aligned}R_n^{\text{swa}}(T,\mathcal{F}) &= \max_{F\in\mathcal{F}}\sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t=a]y_a^t - \sum_{t=1}^{T}\sum_{a\in A_n}\mathbf{1}[a_n^t=a]y_{F(a)}^t \\ &= \max_{F\in\mathcal{F}}\sum_{a\in A_n} L_a^T - \sum_{a\in A_n}\tilde{L}_{a,F(a)}^T = \underbrace{\sum_{a\in A_n}(L_a^T-\hat{L}_a^T)}_{=:(\text{a})} + \underbrace{\sum_{a\in A_n}(\hat{L}_a^T-\hat{L}_{a,F(a)}^T)}_{=:(\text{b})} + \underbrace{\sum_{a\in A_n}(\hat{L}_{a,F(a)}^T-\tilde{L}_{a,F(a)}^T)}_{=:(c)},\end{aligned} \tag{14}$$

where the second equality is due to (13) and the definition of $\tilde{L}_{a,F(a)}^T := \sum_{t=1}^{T}\mathbf{1}[a_n^t=a]y_{F(a)}^t$.

We first show how to bound (a). By definition of L_a^T and $\hat{L}_a^T$, we have that

$$L_a^T-\hat{L}_a^T = \sum_{t=1}^{T}\sum_{a'\in A_n}Y_{a,a'}^t - \sum_{t=1}^{T}\sum_{a'\in A_n}q_{a,a'}^t\hat{Y}_{a,a'}^t = \sum_{t=1}^{T}\sum_{a'\in A_n}Y_{a,a'}^t\left(1-\frac{q_{a,a'}^t}{q_{a,a'}^t+\gamma_t}\right) = \sum_{t=1}^{T}\gamma_t\sum_{a'\in A_n}\hat{Y}_{a,a'}^t.$$

Thus, (a) is bounded by $\sum_{t=1}^{T}\gamma_t\sum_{a\in A_n}\sum_{a'\in A_n}\hat{Y}_{a,a'}^t$.

Then, we show how to bound (b). Let $W_n^t := \prod_{a\in A_n}\sum_{a'\in A_n}\exp(-\eta_{t+1}\hat{L}_{a,a'}^t)$, and we have that $W_n^0 = \prod_{a\in A_n}\sum_{a'\in A_n}\exp(0) = (K_n)^{K_n}$. Note that $W_n^T = W_n^0\frac{W_n^1}{W_n^0}\cdots\frac{W_n^T}{W_n^{T-1}} = (K_n)^{K_n}\prod_{t=1}^{T}\frac{W_n^t}{W_n^{t-1}}$. Then we have

$$\exp(-\sum_{a\in A_n}\eta_{T+1}\hat{L}_{a,F(a)}^T) = \prod_{a\in A_n}\exp(-\eta_{T+1}\hat{L}_{a,F(a)}^T) \leq \prod_{a\in A_n}\sum_{a'\in A_n}\exp(-\eta_{T+1}\hat{L}_{a,a'}^T) = (K_n)^{K_n}\prod_{t=1}^{T}\frac{W_n^t}{W_n^{t-1}}, \tag{15}$$

where the inequality is due to that $\exp\left(-\eta_T \hat{L}^T_{w,w'}\right) \geq 0$. Then, by the definition of $q^t_{w,w'}$ in (5), we obtain that

$$
\begin{aligned}
\frac{W_n^t}{W_n^{t-1}} &= \frac{\prod\limits_{a \in A_n} \sum\limits_{a' \in A_n} \exp\left(-\eta_t \hat{L}^{t-1}_{a,a'}\right) \exp\left(-\eta_t \hat{Y}^t_{a,a'}\right)}{\prod\limits_{a \in A_n} \sum\limits_{a' \in A_n} \exp\left(-\eta_t \hat{L}^{t-1}_{a,a'}\right)} \\
&= \prod_{a \in A_n} \sum_{a' \in A_n} \frac{\exp\left(-\eta_t \hat{L}^{t-1}_{a,a'}\right)}{\sum\limits_{a' \in A_n} \exp\left(-\eta_t \hat{L}^{t-1}_{a,a'}\right)} \exp\left(-\eta_t \hat{Y}^t_{a,a'}\right) \\
&= \prod_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \exp\left(-\eta_t \hat{Y}^t_{a,a'}\right) \leq \prod_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \exp\left(-\eta_T \hat{Y}^t_{a,a'}\right) \\
&\leq \prod_{a \in A_n} \left(\sum_{a' \in A_n} q^t_{a,a'} - \eta_T \sum_{a' \in A_n} q^t_{a,a'} \hat{Y}^t_{a,a'} + \frac{\eta_T^2}{2} \sum_{a' \in A_n} q^t_{a,a'} \left(\hat{Y}^t_{a,a'}\right)^2 \right) \\
&\leq \prod_{a \in A_n} \exp\left(-\eta_T \sum_{a' \in A_n} q^t_{a,a'} \hat{Y}^t_{a,a'} + \frac{\eta_T^2}{2} \sum_{a' \in A_n} q^t_{a,a'} \left(\hat{Y}^t_{a,a'}\right)^2 \right) \\
&= \exp\left(-\eta_T \sum_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \hat{Y}^t_{a,a'} + \frac{\eta_T^2}{2} \sum_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \left(\hat{Y}^t_{a,a'}\right)^2 \right),
\end{aligned}
\tag{16}
$$

where the first inequality is due to that η_t is a non-increasing parameter, the second inequality is due to that $\exp(x) \leq 1 + x + \frac{x^2}{2}$ for any $x \leq 0$, and the third inequality is due to that $1 + x \leq \exp(x)$ for any $x \in \mathbb{R}$. Combining (16) and (15), and taking the logarithm for both sides of the above inequality, we have that

$$
-\sum_{a \in A_n} \eta_T \hat{L}^T_{a,F(a)} \leq K_n \log(K_n) - \sum_{a \in A_n} \eta_T \underbrace{\sum_{t=1}^T \sum_{a' \in A_n} q^t_{a,a'} \hat{Y}^t_{a,a'}}_{=: \hat{L}^T_a \text{ (by definition of } \hat{L}^T_a)} + \frac{\eta_T^2}{2} \sum_{t=1}^T \sum_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \left(\hat{Y}^t_{a,a'}\right)^2 .
$$

Dividing both sides by $\eta_T > 0$, with rearrangement, we have

$$
\begin{aligned}
\sum_{a \in A_n} \hat{L}^T_a - \sum_{a \in A_n} \hat{L}^T_{a,F(a)} &\leq \frac{K_n \log(K_n)}{\eta_T} + \frac{\eta_T}{2} \sum_{t=1}^T \sum_{a \in A_n} \sum_{a' \in A_n} q^t_{a,a'} \left(\hat{Y}^t_{a,a'}\right)^2 \\
&\leq \frac{K_n \log(K_n)}{\eta_T} + \sum_{t=1}^T \frac{\eta_t}{2} \sum_{a \in A_n} \sum_{a' \in A_n} \hat{Y}^t_{a,a'},
\end{aligned}
\tag{17}
$$

where the second inequality is due to that η_t is a non-increasing parameter and the fact that $q^t_{a,a'} \hat{Y}^t_{a,a'} \leq 1$. Combining with the bound of (a), we have

$$
\sum_{a \in A_n} \left(L^T_a - \tilde{L}^T_{a,F(a)} \right) \leq \frac{K_n \log(K_n)}{\eta_T} + \sum_{t=1}^T \left(\frac{\eta_t}{2} + \gamma_t \right) \sum_{a \in A_n} \sum_{a' \in A_n} \hat{Y}^t_{a,a'} + \sum_{a \in A_n} \left(\hat{L}^T_{a,F(a)} - \tilde{L}_{a,F(a)} \right).
$$

Let $\gamma_t = \eta_t / 2$. By invoking Lemma 5.2, with probability at least $1 - \delta$, we have the following inequality held:

$$
\begin{aligned}
\sum_{a \in A_n} \left(L^t_a - \tilde{L}^T_{a,a'} \right) &\leq \frac{K_n \log(K_n)}{\eta_T} + \sum_{t=1}^T \eta_t \left(\sum_{a \in A_n} \sum_{a' \in A_n} \tilde{Y}^t_{a,a'} \right) + 2 \log\left(\frac{2}{\delta}\right) + \frac{1}{\eta_T} \log\left(\frac{2K_n^{K_n}}{\delta}\right) \\
&\leq \frac{K_n \log(K_n) + K_n \log(2K_n/\delta)}{\eta_T} + \sum_{t=1}^T \eta_t K_n + 2 \log\left(\frac{2}{\delta}\right),
\end{aligned}
$$

where the last inequality is due to that $\sum\limits_{a \in A_n} \sum\limits_{a' \in A_n} \tilde{Y}^t_{a,a'} = \sum\limits_{a \in A_n} \sum\limits_{a' \in A_n} \mathbf{1}[a^t_n = a] y^t_{a'} \leq K_n$ and $\log\left(\frac{2K_n^{K_n}}{\delta}\right) \leq K_n \log(2K_n/\delta)$ for $\delta \in (0, 1)$.

Letting $\eta_t = \sqrt{\frac{\log(K_n)}{t}}$, we have

$$R_n^T(T,\mathcal{F}) \leq 2K_n\sqrt{T\log(K_n)} + K_n\sqrt{\log(K_n)}\sum_{t=1}^{T}\sqrt{\frac{1}{t}} + \left(2 + K_n\sqrt{\frac{T}{\log K_n}}\right)\log(\frac{2}{\delta}).$$

When $\eta_t = \sqrt{\frac{\log(K_n)+\log(2K_n/\delta)}{t}}$, the above inequality becomes

$$R_n^T(T,\mathcal{F}) \leq K_n\sqrt{T(\log(K_n)+\log(2K_n/\delta))} + K_n\sqrt{(\log(K_n)+\log(2K_n/\delta)}\sum_{t=1}^{T}\frac{1}{t} + 2\log(\frac{2}{\delta}).$$

Theorem 5.3 follows by $\sum\limits_{t=1}^{T}\sqrt{\frac{1}{t}} \leq 2\sqrt{T}$. □