Multi-agent Bandits in an Unknown General-sum Game Environment

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

Abstract

In this report, we study a distributed multi-armed bandit problem involving a set of agents $\mathcal{N}$ in an unknown general-sum game. In each round, each agent $n \in \mathcal{N}$ needs to play an arm from a (possibly different) arm set with K_n arms, and receives the reward of the played arm that is affected by other agents' actions. The objective of each agent is to accumulate as many rewards as possible within T rounds. The new challenges come from the unknown general-sum game setting, i.e., each agent has no knowledge about the environment, e.g., the number of other agents and the arms played by other agents. This problem is motivated by real-world applications such as congestion control, medium access, and routing in computer networks, as well as in other domains. To address the problem, we proposed an exponential-weighting-based algorithm called *learning for correlated equilibrium (LCE)* and proved that LCE can maximize the cumulative rewards for each agent with a performance loss bounded by $O(\sqrt{TK_n \log(K_n)})$ comparing with always playing the best arm in hindsight, which has shaved an extra factor of K_n when compared with the best result so far. It is also guaranteed that an ϵ-correlated equilibrium for unknown general-sum games can be achieved in a polynomial number of rounds if LCE is played by all agents. Furthermore, we conducted dynamic distributed medium access experiments to verify the effectiveness of LCE.

1. Model and Problem Formulation

1.1. The MAB-UG Model

We study a multi-agent bandit problem in an unknown general-sum game environment (MAB-UG), where the reward of each agent's decision will be affected by that 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, and the rewards of the arms that are not played.

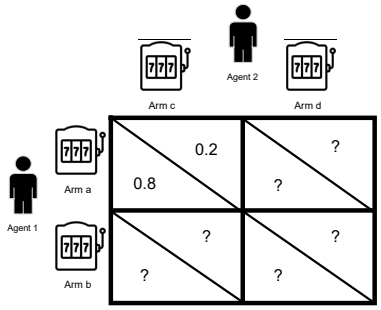

Figure 1. An example of MAB-UG with two agents and two arms for each agent.

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 $a_n \in A_n$ is determined by function $u_n : \mathcal{A} \to [0, 1]$, which maps the actions of all agents to agent n's rewards $u_n(a_n; \mathbb{A}_{-n})$. [1] Furthermore, let $\mathcal{U} := \{u_1, \ldots, u_N\}$ be the set of reward functions. Note that neither $\mathcal{N}$ or $\mathcal{U}$ 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

[1]Department of Computer Science, University of Victoria, BC, Canada. Correspondence to: Zhiming Huang <zhiminghuang@uvic.ca>, Jianping Pan <pan@uvic.ca>.

[1] $(a_n; \mathbb{A}_{-n})$ is an abbreviation of $\mathbb{A} := (a_1, \ldots, a_n, \ldots, a_N)$ with a highlight of agent n's action a_n against other agents' actions.

agent n can only observe her own instantaneous reward $X_n^t := u_n(a_n^t; \mathbb{A}_{-n}^t)$. [2] Both the actions and the number of other agents cannot be observed. The objective of each agent is to accumulate as many cumulative rewards as possible over a given time horizon.

1.2. 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 & 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 using a learning algorithm compares the cumulative expected reward of the learning algorithm with that of the optimal arm in hindsight up to round T, which is defined as follows:

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

where the expectation is taken with respect to the randomness of the actions for all agents. 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:

$$R_n^{\text{int}}(T) := \max_{a, a' \in A_n} \mathbf{E}\left[\sum_{t=1}^{T} r_{(a,a'),n}^t\right],$$

where

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

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

Definition 1.1. 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

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

[2] For the convenience of algorithm description and analysis, we sometimes use an equivalent notion called the instantaneous loss by $1 - X_n^t$.

When $\epsilon = 0$, then $\mathbf{P}$ is the correlated equilibrium, which is more general than the well-known Nash equilibrium, as the correlated equilibrium does not require the independence among actions. To give an intuition about the ϵ-correlated equilibrium, consider a case in congestion control where a *mediator* (e.g., a router or switch) draws an action profile from $\mathbf{P}$ and privately recommends each action (i.e., the congestion window size) to the corresponding host. If no host has an incentive more than ϵ to choose a different congestion window, provided that other hosts follow the router'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.

To get tight bounds for both the external regret and internal regret, we consider a more general notion of regret, called the *swap regret* (Blum & Mansour, 2007), which can unify both the external regret and internal regret into the same framework by a swap function $F_n : A_n \to A_n$ 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 swap regret for agent n with $\mathcal{F}$ up to round T is defined as follows:

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

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 $a \in A_n$, $F_a \in \mathcal{F} : A_n \to a$. 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 for a general $\mathcal{F}$, we can minimize the internal and external regrets at the same time, and achieve the ϵ-correlated equilibrium for all agents.

Therefore, we convert the problem of maximizing the cumulative rewards to the problem of minimizing the swap regret for each agent. The challenges to address the problem come from the facts that each agent has no prior knowledge of the environment and each agent can only observe the rewards of their own played arms. Thus, each agent faces a dilemma between the exploration and exploitation. The exploration means each agent needs to play each arm for more information, and the exploitation means that each agent needs to play the currently-known best arm to gain more rewards. It is also a challenge to design a swap-regret-minimizing algorithm, as currently most bandit algorithms are designed for the minimization of the external regret (Lattimore & Szepesvári, 2020).

2. The LCE Algorithm

To tackle the challenges of balancing the tradeoff between the exploration and exploitation, we adopt the exponen-

tial weighting techniques (Auer et al., 2002). The main idea is to play an arm sampled from a probability distribution which gives more weights to the arms that have returned more rewards in history. Such a technique has been largely used in the adversarial bandit problems where the rewards of arms are random without stochastic assumptions. Furthermore, we also incorporate the variance-reduction technique (Kocák et al., 2014) to ensure that the proposed algorithm can achieve a sublinear regret over time without a large variance. However, trivial application of such techniques does not work for MAC-UG, as they cannot guarantee that each agent can reach the desired equilibrium.

To address the challenges of minimizing the swap regret, we use the concept of meta-distributions introduced in Blum & Mansour (2007). Note that although we adopt the similar idea for algorithm design, we give a different analysis to obtain a tighter regret bound, and we note that the difference is not trivial, as discussed in the proof sketch for Theorem 3.1. 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 $\sum_{a' \in A_n} q_{a,a'}^t = 1$, where $q_{a,a'}^t$ can be understood as the probability of playing arm a' instead of playing arm a. Denote by $\mathbb{Q}_n^t := [Q_a^t]_{a \in A_n}$ the $N_k \times N_k$ matrix with each row being Q_a^t. Then, we determine the sample distribution P_n^t by solving the following equations:

$$P_n^t = P_n^t \mathbb{Q}_n^t, \tag{2}$$

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 (2) 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 choosing any arm $a \in A_n$ and then playing a' according to Q_a^t.

By carefully integrating all the above techniques, we design the *learning for correlated equilibrium (LCE)* algorithm as shown in Alg. 1. At the very beginning, LCE sets all the meta-distributions to be uniform with maximal uncertainty by letting $q_{a,a'}^1 = \frac{1}{K_n}$ and defines $L_{a,a'}^t$ the cumulative loss by the end of t for each action-pair $a, a' \in A_n$ with an initial value $L_{a,a'}^0 = 0$, as shown in Line 3.

In each round $t = 1, \ldots, T$, LCE first calculates the probability distribution P_n^t based on (2), according to which arm a_n^t is played, and observes reward X_n^t for that arm, as shown in Lines 6 to 8. Then, LCE needs to update each meta-distribution $Q_a^{t+1}, \forall a \in A_n$ based on the observed reward X_n^t, as shown in Lines 10 to 14. Each meta-distribution receives the observed reward in the form of the divided loss $Y_{a,a'}^t$ according to the contributions to play arm a_n^t. Then,

Algorithm 1 The LCE algorithm

1: **Input:** n, T, A_n, η
2: // Initialization
3: Set $q_{a,a'}^1 = \frac{1}{K_n}$ and $\hat{L}_{a,a'}^0 = 0, \forall a, a' \in A_n$
4: **for** $t = 1, \ldots, T$ **do**
5: // Compute the sample distribution, play arms and observe rewards
6: Calculate P_n^t based on (2)
7: Play an arm $a_n^t \sim P_n^t$
8: Observe reward X_n^t
9: // Update each meta-distribution
10: **for** $a \in A_a$ **do**
11: $Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a'] p_a^t q_{a,a'}^t}{p_{a'}^t}(1 - X_n^t), \forall a' \in A_n$
12: $\hat{L}_{a,a'}^t = \hat{L}_{a,a'}^{t-1} + \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma}, \forall a' \in A_n$
13: Calculate Q_a^{t+1} based on (5)
14: **end for**
15: **end for**

the divided loss $Y_{a,a'}^t$ is defined as follows:

$$Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a'] p_a^t q_{a,a'}^t}{p_{a'}^t}(1 - X_n^t), \forall a' \in A_n. \tag{3}$$

To see the reason for (3), we can add up the divided loss for all meta-distributions, which gives

$$\sum_{a \in A_n} Y_{a,a_n^t}^t = \frac{\sum_{a \in A_n} p_a^t q_{a,a_n^t}^t}{p_{a_n^t}^t}(1 - X_n^t) = \frac{p_{a_n^t}^t}{p_{a_n^t}^t}(1 - X_n^t) = (1 - X_n^t).$$

This means the sum of divided loss suffered by each meta-distribution equals the loss suffered by LCE. Note that as X_n^t ranges in $[0,1]$, $Y_{a,a'}^t$ also ranges in $[0,1]$.

Each meta-distribution then uses a biased loss estimator with the bias parameter γ defined as follows [3]

$$\hat{Y}_{a,a'}^t := \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma}. \tag{4}$$

The cumulative estimated loss $\hat{L}_{a,a'}^t = \hat{L}_{a,a'}^{t-1} + \hat{Y}_{a,a'}^t$ for meta-distribution Q_a^t can be updated, as shown in Line 12. At the end of round t, we can use the exponential weighting techniques to update the meta-distributions by

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

where η is a parameter controlling the learning rate and its value will be given in Theorem 3.1.

[3] γ is used for smoothing the meta-distributions so that arms with large losses in the past can still be chosen occasionally for exploration. A possible value of γ is given in Theorem 3.1.

3. Analytical Results

3.1. Regret Bound

The swap regret defined in (1) for each agent playing Alg. 1 is bounded by the following theorem:

Theorem 3.1. *If LCE is run with* $\eta = \sqrt{\frac{\log(K_n)}{T}}$ *and* $\gamma = \eta/2$*, we have*

$$R_n^{\text{swa}}(T) \leq 2\sqrt{TK_n \log(K_n)} + 2K_n(\log(K_n) + 1).$$

Proof Sketch. In Blum & Mansour (2007), each meta-distribution is generated by a multi-armed bandit algorithm, and each bandit algorithm is treated as a black-box in the algorithm analysis. The swap regret bound is decomposed to be the sum of the expected external regret of K_n bandit algorithms, each with an expected external regret of $O(\sqrt{TK_n \log K_n})$ (Auer et al., 2002), which results in the swap regret bound of $O(K_n\sqrt{TK_n \log K_n})$.

Different from Blum & Mansour (2007), we open the black-box by first using a martingale-based technique to derive a high-probability regret bound for any actual sequence of instantaneous losses, and then integrating the tails of the high-probability bound to obtain the swap regret. The benefit of such a technique can be twofold. First, as the instantaneous loss suffered by each agent in each round is dependent on other agents' actions, and the actions of all agents are dependent on their previous actions and losses, the expectation of $Y_{a,a'}^t$ with respect to all the randomness is very difficult to compute. Therefore, such a technique can avoid these difficulties. Second, when we consider the effect from all meta-distributions, such a technique can utilize an important property: $\sum_{a\in A_n}\sum_{a'\in A_n} Y_{a,a'}^t \leq \sum_{a'\in A_n}\sum_{a\in A_n} \frac{p_a^t q_{a,a'}^t \mathbf{1}[a_n^t=a']}{p_{a'}^t} \leq \sum_{a'\in A_n} \mathbf{1}[a_n^t = a'] = 1$, which can be ignored if taking the expectation of each meta-algorithm first and then doing the summation.

The proof starts with a key step to convert the swap regret bound of the whole algorithm to the bound of each meta-distribution as follows:

$$R_n^{\text{swa}} \leq \mathbf{E}\left[\sum_{a\in A_n}\left(\tilde{L}_a^T - \hat{L}_{a,a'}^T\right)\right],$$

where $\tilde{L}_a^T := \sum_{t=1}^T \sum_{a'\in A_n} Y_{a,a'}^t$ is the actual cumulative loss suffered by meta-distribution Q_a^t over T rounds. Then, we derive high-probability bounds for $\tilde{L}_a^T - \hat{L}_{a,a'}^T$ with any random actual sequence of $Y_{a,a'}, \forall a, a' \in A_n$, and integrate the tail for the high-probability bounds to obtain the expectation. Decompose $\tilde{L}_a^T - \hat{L}_{a,a'}^T$ as follows:

$$\tilde{L}_a^T - \hat{L}_{a,a'}^T = \underbrace{(\tilde{L}_a^T - \hat{L}_a^T)}_{=:(\text{a})} + \underbrace{(\hat{L}_a^T - \hat{L}_{a,a'}^T)}_{=:(\text{b})},$$

where $\hat{L}_a^T := \sum_{t=1}^T q_{a,a'}^t \hat{Y}_{a,a'}^t$. We prove that Term (a) can be bounded by $\gamma \sum_{a'\in A_n} \hat{L}_{a,a'}^T$ and Term (b) can be bounded with the standard analysis for the exponential weighting techniques by $\frac{\log(K_n)}{\eta} + \frac{\eta}{2}\sum_{a'\in A_n} \hat{L}_{a,a'}^T$ (see Lemma A.1 in Appendix A). By using a martingale-based analysis, we give a high-probability tight bound for $\hat{L}_{a,a'}^t$, as we proved in Lemma A.2 in Appendix B. Combining (a) and (b) together and invoking Lemma A.2, with probability at least $1-\delta$, where $\delta \in (0,1)$, we obtain that:

$$\sum_{a\in A_n}\left(\tilde{L}_a^T - \hat{L}_{a,a'}^T\right) \leq \frac{K_n \log(K_n)}{\eta} + \left(\frac{\eta}{2}+\gamma\right)T + \left(\frac{\eta}{2}+\gamma\right)K_n \frac{\log\left(\frac{K_n}{\delta}\right)}{\gamma}.$$

The theorem follows by substituting $\eta = \sqrt{\frac{K_n \log(K_n)}{T}}$ and $\gamma = \eta/2$, and integrating the tail. The detailed proof can be found in Appendix A. □

When $T > K_n \log(K_n)$, the bound in Theorem 3.1 can be written as $O(\sqrt{TK_n \log(K_n)})$. To the best of our knowledge, currently our bound is the best for the swap regret, which shaves off the extra K_n factors from the upper swap regret bound for adversarial bandits proved in Blum & Mansour (2007). It also matches the best external regret bound for the online learning algorithms based on the exponential weighting techniques (Auer et al., 2002). However, our swap regret bound can be further improved, as the lower swap regret bound is $\Omega(\sqrt{TN})$ as shown in Blum & Mansour (2007)

3.2. Convergence to Correlated Equilibrium

Denote by $\hat{\mathbf{P}}^T(\mathbb{A}) := \frac{1}{T}\sum_{t=1}^T \mathbf{1}\left[\mathbb{A}_t = \mathbb{A}\right], \mathbb{A} \in \mathcal{A}$ the empirical distribution of the joint actions played by all agents.

Theorem 3.2. *If every agent* $n \in \mathcal{N}$ *plays the LCE 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.*

Proof Sketch. The result directly follows Theorem 3 in Blum & Mansour (2007). As Blum & Mansour (2007) does not give a proof, we provide a proof sketch here to give an intuition, and more detailed proofs can be found in Sec. 7.4 in Cesa-Bianchi & Lugosi (2006). As explained in Sec. 1, the swap regret can boil down to the internal regret, which means LCE also has a bounded internal regret for each agent n and any action pairs $a, a' \in A_n$:

$$\mathbf{E}\left[\sum_{t=1}^T r_{(a,a'),n}^t\right] = \mathbf{E}\left[\sum_{t=1}^T \mathbf{1}\left[a_n^t = a\right]\left(u_n\left(a'; \mathbb{A}_{-n}^t\right) - u_n\left(a; \mathbb{A}_{-n}^t\right)\right)\right] = O(\sqrt{TK_n \log K_n}).$$

Observe that $r_{(a,a'),n}^t - \mathbf{E}\left[r_{(a,a'),n}^t\right]$ is a martingale difference sequence, and by applying the Hoeffding-Azuma

inequality and the union bound, with the probability at least $1-\delta$ for all N agents, we have for any fixed $a, a' \in A_n$:

$$\sum_{t=1}^{T} r^t_{(a,a'),n} \leq \sum_{t=1}^{T} \mathbf{E}\left[r^t_{(a,a'),n}\right] + \sqrt{\frac{T\log(\frac{N}{\delta})}{2}} = O(\sqrt{TK_n\log(K_nN/\delta)}).$$

The theorem follows by dividing both sides by T:

$$\sum_{\mathbb{A}:a_n=a} \hat{\mathbf{P}}(\mathbb{A})\left(u_n\left(a';\mathbb{A}_{-n}\right)-u_n\left(\mathbb{A}\right)\right) = \frac{1}{T}\sum_{t=1}^{T} r^t_{(a,a'),n} = O(\sqrt{\frac{K_n\log(K_nN/\delta)}{T}}),$$

which coincides with the definition of the ϵ-correlated equilibrium in Sec. 1. □

The above theorem implies that if all agents play the LCE algorithm for a long time, then all agents can reach the correlated equilibrium ($\epsilon = 0$). Furthermore, Theorem 3.2 also implies that LCE can be used to find an ϵ-correlated equilibrium efficiently, as shown in the following corollary.

Corollary 3.3. *The LCE algorithm can find an ϵ-correlated equilibrium for unknown general-sum games in $O(\max_{n\in\mathcal{N}} \frac{K_n\log(K_nN/\delta)}{\epsilon^2})$ rounds with probability $1-\delta$.*

Proof. Assume we find an ϵ-correlated equilibrium for a game with N agents in T rounds. Then, we have

$$\epsilon T = O(\max_{n\in\mathcal{N}} \sqrt{TK_n\log(K_nN/\delta)}),$$

which gives $T = O(\max_{n\in\mathcal{N}} \frac{K_n\log(K_nN/\delta)}{\epsilon^2})$. □

The above corollary shows that an ϵ-correlated equilibrium can be achieved by the LCE algorithm in a polynomial number of rounds.

3.3. Time and Space Complexity

In each round, each agent needs to first calculate P^t_n based on (2), which can be regarded as the calculation of a stationary distribution for the Markov process defined by Q^t_n, and can be achieved within $O(K_n^2)$ for K_n states (Feinberg & Chiu, 1987). Then, each meta-distribution needs $O(K_n)$ time to be updated for K_n arms. Therefore, the time complexity for LCE 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. Therefore, the space complexity for LCE is $O(K_n^2)$.

4. Conclusion

In this report, we have proposed the LCE algorithm to address the MAB-UG problem with a provable near-optimal swap regret bound of $\sqrt{TK_n\log(K_n)}$ for K_n arms in T rounds, which shaves off K_n from the existing work (Blum & Mansour, 2007). In addition, if every agent plays LCE, then the empirical distribution of joint plays will converge to an ϵ-correlated equilibrium. An experiment has been given to verify the performance of LCE.

Nevertheless, the time and space consumption of the LCE algorithm requires $O(K_n^2)$, and the number of the arms is discrete and finite, which needs further improvement.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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

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

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

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

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

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

A. Proof of Theorem 3.1

We start by introducing the notations and lemmas that will be used in the proof of Theorem 3.1. The proofs of all the lemmas can be found in Appendix B.

Let $x_a^t := u_n(a; \mathbb{A}_{-n}^t)$ and $y_a^t := 1 - x_a^t$. Recall $Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a']p_a^t q_{a,a'}^t y_{a'}^t}{p_{a'}^t}$ is the loss observed by meta-distribution Q_a^t, and $\hat{Y}_{a,a'}^t = \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma}$ is the loss estimator of a' for meta-distribution Q_a^t. Then, we further define some notations as follows. Denote by $\hat{L}_a^T := \sum_{t=1}^T \sum_{a' \in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t$ and $\tilde{L}_a^T := \sum_{t=1}^T \sum_{a' \in A_n} Y_{a,a'}^t$ the cumulative estimated and actual loss by meta-distribution Q_a^t over T rounds, respectively. By the relationship between P_n^t and Q_a^t, we have

$$\sum_{a \in A_n} \tilde{L}_a^T = \sum_{t=1}^T \sum_{a' \in A_n} \sum_{a \in A_n} \frac{\mathbf{1}[a_n^t = a']p_a^t q_{a,a'}^t}{p_{a'}^t} y_{a'}^t = \sum_{t=1}^T \sum_{a' \in A_n} \mathbf{1}[a_n^t = a']y_{a'}^t = \sum_{t=1}^T \sum_{a \in A_n} \mathbf{1}[a_n^t = a]y_a^t, \tag{6}$$

where the second inequality is due to that $\sum_{a \in A_n} p_a^t q_{a,a'}^t = p_{a'}^t$. Furthermore, recall $\hat{L}_{a,a'}^T := \sum_{t=1}^T \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma}$ and let $\mathbf{E}_t[\cdot] := \mathbf{E}[\cdot | a_n^1, X_n^1, \ldots, a_n^t, X_n^t]$ be the expectation conditioned on the previous actions and rewards up to round t. Then, by using the law of total expectation, we have the following equation held for any $a' \in A_n$:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{a \in A_n} \hat{L}_{a,a'}^T\right] \leq \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \frac{Y_{a,a'}^t}{q_{a,a'}^t}\right] = \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \frac{\mathbf{1}[a_n^t = a']p_a^t y_{a'}^t}{p_{a'}^t}\right] = \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \mathbf{E}_{t-1}\left[\frac{\mathbf{1}[a_n^t = a']p_a^t y_{a'}^t}{p_{a'}^t}\right]\right] \\
&= \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \frac{p_{a'}^t}{p_{a'}^t} \mathbf{E}_{t-1}\left[p_a^t y_{a'}^t\right]\right] = \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \mathbf{E}_{t-1}\left[\mathbf{1}[a_n^t = a]y_{a'}^t\right]\right] = \mathbf{E}\left[\sum_{a \in A_n} \sum_{t=1}^T \mathbf{1}[a_n^t = a]y_{a'}^t\right].
\end{aligned} \tag{7}$$

Lemma A.1. *For any $a \in A_n$, and $\eta > 0$, we have*

$$\hat{L}_a^T - \hat{L}_{a,a'}^T \leq \frac{\log(K_n)}{\eta} + \frac{\eta}{2} \sum_{a' \in A_n} \hat{L}_{a,a'}^T.$$

Denote by $L_{a,a'}^T := \sum_{t=1}^T Y_{a,a'}^t$, we have the following inequality held:

Lemma A.2. *Let $\delta \in (0,1)$. With probability at least $1 - \delta$, we have*

$$\sum_{a' \in A_n} \left(\hat{L}_{a,a'}^T - L_{a,a'}^T\right) < \frac{\log\left(\frac{1}{\delta}\right)}{\gamma}.$$

With the above notations and lemmas, we give the proof of Theorem 3.1 as follows:

Proof of Theorem 3.1. The swap regret defined in (1) can be rewritten in the loss form:

$$R_n^{\text{swa}}(T, \mathcal{F}) = \max_{F \in \mathcal{F}} \mathbf{E}\left[\sum_{t=1}^T \sum_{a \in A_n} \mathbf{1}[a_n^t = a]y_a^t - \sum_{t=1}^T \sum_{a \in A_n} \mathbf{1}[a_n^t = a]y_{F(a)}^t\right]$$

With (6) and (7), the swap regret can be bounded as follows:

$$\begin{aligned}
&\max_{F \in \mathcal{F}} \mathbf{E}\left[\sum_{t=1}^T \sum_{a \in A_n} \mathbf{1}[a_n^t = a]y_a^t - \sum_{t=1}^T \sum_{a \in A_n} \mathbf{1}[a_n^t = a]y_{F(a)}^t\right] \\
&\leq \max_{F \in \mathcal{F}} \mathbf{E}\left[\sum_{a \in A_n} \tilde{L}_a^T - \sum_{a \in A_n} \hat{L}_{a,F(a)}^T\right] = \max_{F \in \mathcal{F}} \mathbf{E}\left[\sum_{a \in A_n} \left(\tilde{L}_a^T - \hat{L}_{a,F(a)}^T\right)\right].
\end{aligned}$$

Thus, to bound $R_n^{\text{swa}}(T, \mathcal{F})$, it is sufficient to bound $\tilde{L}_a^T - \hat{L}_{a,a'}^T$ for any $a' \in A_n$ and for all meta-distributions. The above equation is the key step to bound the swap regret, as now we have converted the bound of swap regret to the bound of $\tilde{L}_a^T - \hat{L}_{a,a'}^T$ for each meta-distribution Q_a^t over T rounds. Decompose $\tilde{L}_a^T - \hat{L}_{a,a'}^T$ as follows:

$$\tilde{L}_a^T - \hat{L}_{a,a'}^T = \underbrace{(\tilde{L}_a^T - \hat{L}_a^T)}_{=:(\text{a})} + \underbrace{(\hat{L}_a^T - \hat{L}_{a,a'}^T)}_{=:(\text{b})}.$$

Term (a) can be bounded as follows:

$$\begin{aligned}
\tilde{L}_a^T - \hat{L}_a^T &= \sum_{t=1}^T \sum_{a' \in A_n} Y_{a,a'}^t - \sum_{t=1}^T \sum_{a' \in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t \\
&= \sum_{t=1}^T \sum_{a' \in A_n} Y_{a,a'}^t \left(1 - \frac{q_{a,a'}^t}{q_{a,a'}^t + \gamma}\right) = \gamma \sum_{t=1}^T \sum_{a' \in A_n} \hat{Y}_{a,a'}^t = \gamma \sum_{a' \in A_n} \hat{L}_{a,a'}^T.
\end{aligned}$$

Term (b) can be bounded by invoking Lemma A.1, and combining with the bound of Term (a), we have

$$\tilde{L}_a^T - \hat{L}_{a,a'}^T \le \frac{\log(K_n)}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{a' \in A_n} \hat{L}_{a,a'}^T.$$

Let $\frac{\delta}{K_n} \in (0, 1)$. By invoking Lemma A.2 and the union bound, with probability at least $1 - \delta$, we have the following inequality held:

$$\begin{aligned}
&\sum_{a \in A_n} \left(\tilde{L}_a^T - \hat{L}_{a,a'}^T\right) \le \sum_{a \in A_n} \left(\frac{\log(K_n)}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \left(\sum_{a' \in A_n} L_{a,a'}^T + \frac{\log\left(\frac{K_n}{\delta}\right)}{\gamma}\right)\right) \\
&\le \frac{K_n \log(K_n)}{\eta} + \left(\frac{\eta}{2} + \gamma\right) T + \left(\frac{\eta}{2} + \gamma\right) K_n \frac{\log\left(\frac{K_n}{\delta}\right)}{\gamma},
\end{aligned}$$

where the last inequality is due to $\sum_{a \in A_n} \sum_{a' \in A_n} L_{a,a'}^T = \sum_{a' \in A_n} \sum_{t=1}^T \sum_{a \in A_n} \frac{p_a^t q_{a,a'}^t \mathbf{1}[a_n^t = a'] y_{a'}^t}{p_{a'}^t} \le \sum_{a' \in A_n} \sum_{t=1}^T \mathbf{1}[a_n^t = a'] = T$. Therefore, by letting $\gamma = \frac{\eta}{2}$ and $\eta = \sqrt{\frac{K_n \log(K_n)}{T}}$, we have with probability at least $1 - \delta$:

$$\sum_{a \in A_n} \left(\tilde{L}_a^T - \hat{L}_{a,a'}^T\right) \le 2\sqrt{T K_n \log(K_n)} + 2K_n \log\left(\frac{K_n}{\delta}\right). \tag{8}$$

Theorem 3.1 follows by integrating the tail

$$\mathbf{E}[W] \le \int_0^1 \frac{1}{\delta} \mathbb{P}\left(W > \log \frac{1}{\delta}\right) d\delta, \tag{9}$$

where

$$W := \frac{1}{2K_n} \sum_{a \in A_n} \left(\tilde{L}_a^T - \hat{L}_{a,a'}^T\right) - \left(\sqrt{K_n T \log(K_n)} + \log(K_n)\right).$$

□

B. Proof of Lemmas

Proof of Lemma A.1. Let $W_n^t := \sum_{a' \in A_n} \exp\left(-\eta \hat{L}_{a,a'}^t\right)$ and thus $W_n^0 = \sum_{a' \in A_n} \exp\left(-\eta \hat{L}_{a,a'}^0\right) = \sum_{a' \in A_n} \exp(0) = K_n$. Note that

$$W_n^T = W_n^0 \frac{W_n^1}{W_n^0} \cdots \frac{W_n^T}{W_n^{T-1}} = K_n \prod_{t=1}^T \frac{W_n^t}{W_n^{t-1}}.$$

Then we have

$$\exp\left(-\eta \hat{L}_{a,a'}^T\right) \leq \sum_{a' \in A_n} \exp\left(-\eta \hat{L}_{a,a'}^T\right) = W_n^T = K_n \prod_{t=1}^T \frac{W_n^t}{W_n^{t-1}}. \tag{10}$$

Recall that $\hat{L}_{a,a'}^t = \hat{L}_{a,a'}^{t-1} + \hat{Y}_{a,a'}^t$ and the definition of $q_{a,a'}^t$ in (5). We obtain that

$$\begin{aligned}
\frac{W_n^t}{W_n^{t-1}} &= \frac{\sum_{a' \in A_n} \exp\left(-\eta \hat{L}_{a,a'}^{t-1}\right) \exp\left(-\eta \hat{Y}_{a,a'}^t\right)}{W_n^{t-1}} = \sum_{a' \in A_n} q_{a,a'}^t \exp\left(-\eta \hat{Y}_{a,a'}^t\right) \\
&\leq 1 - \eta \sum_{a' \in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t + \frac{\eta^2}{2} \sum_{a' \in A_n} q_{a,a'}^t (\hat{Y}_{a,a'}^t)^2 \\
&\leq \exp\left(-\eta \sum_{a' \in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t + \frac{\eta^2}{2} \sum_{a' \in A_n} q_{a,a'}^t (\hat{Y}_{a,a'}^t)^2\right),
\end{aligned} \tag{11}$$

where the first inequality is due to $\exp(x) \leq 1 + x + \frac{1}{2}x^2$ for all $x \leq 0$, and the second inequality is due to $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$. Combining (11) and (10) and taking the logarithm, we have

$$-\eta \hat{L}_{a,a'}^T \leq \log(k_n) - \eta \underbrace{\sum_{t=1}^T \sum_{a' \in A_n} q_{a,a'}^t \hat{Y}_{a,a'}^t}_{=: \hat{L}_a^T} + \frac{\eta^2}{2} \sum_{t=1}^T \sum_{a' \in A_n} q_{a,a'}^t (\hat{Y}_{a,a'}^t)^2.$$

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

$$\hat{L}_a^T - \hat{L}_{a,a'}^t \leq \frac{\log(K_n)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{a' \in A_n} q_{a,a'}^t (\hat{Y}_{a,a'}^t)^2. \tag{12}$$

As $q_{a,a'}^t \hat{Y}_{a,a'}^t = q_{a,a'}^t \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma} \leq Y_{a,a'}^t \leq 1$, (12) can be bounded as follows:

$$\hat{L}_a^T - \hat{L}_{a,a'}^t \leq \frac{\log(K_n)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{a' \in A_n} \hat{Y}_{a,a'}^t = \frac{\log(K_n)}{\eta} + \frac{\eta}{2} \sum_{a' \in A_n} \hat{L}_{a,a'}^T.$$

□

Proof of Lemma A.2. Define the expectation of $\sum_{a' \in A_n} \hat{Y}_{a,a'}^t$ with respect to distribution Q_a^t as follows:

$$\mu_t := \mathbf{E}_{Q_a^t}\left[\sum_{a' \in A_n} \hat{Y}_{a,a'}^t\right] = \sum_{a' \in A_n} \frac{q_{a,a'}^t Y_{a,a'}^t}{q_{a,a'}^t + \gamma}.$$

Now we show that the covariance with respect to Q_a^t between any two $\hat{Y}_{a,a'}^t$'s is 0. For any $a', a'' \in A_n$, by definition of $Y_{a,a'}^t$ in (3), we have that $Y_{a,a'}^t Y_{a,a''}^t = 0$. Thus, the covariance between $\hat{Y}_{a,a'}^t$ and $\hat{Y}_{a,a''}^t$ can be obtained as follows:

$$\begin{aligned}
\mathrm{Cov}(\hat{Y}_{a,a'}^t, \hat{Y}_{a,a''}^t) &= \mathbf{E}_{Q_a^t}\left[(\hat{Y}_{a,a'}^t - \mathbf{E}_{Q_a^t}[\hat{Y}_{a,a'}^t])(\hat{Y}_{a,a''}^t - \mathbf{E}_{Q_a^t}[\hat{Y}_{a,a''}^t])\right] \\
&= \mathbf{E}_{Q_a^t}\left[\left(\hat{Y}_{a,a'}^t - \frac{q_{a,a'}^t Y_{a,a'}^t}{q_{a,a'}^t + \gamma}\right)\left(\hat{Y}_{a,a''}^t - \frac{q_{a,a''}^t Y_{a,a''}^t}{q_{a,a''}^t + \gamma}\right)\right] \\
&= \mathbf{E}_{Q_a^t}\left[\left(\frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma} - \frac{q_{a,a'}^t Y_{a,a'}^t}{q_{a,a'}^t + \gamma}\right)\left(\frac{Y_{a,a''}^t}{q_{a,a''}^t + \gamma} - \frac{q_{a,a''}^t Y_{a,a''}^t}{q_{a,a''}^t + \gamma}\right)\right] \\
&= \mathbf{E}_{Q_a^t}\left[Y_{a,a'}^t Y_{a,a''}^t \left(\frac{1 - q_{a,a'}^t}{q_{a,a'}^t + \gamma}\right)\left(\frac{1 - q_{a,a''}^t}{q_{a,a''}^t + \gamma}\right)\right] = 0.
\end{aligned}$$

Using the same idea, one can easily verify that $\{Y^t_{a,a'}\}_{a'\in A_n}$ are independent by checking $\mathrm{Cov}\left(\{Y^t_{a,a'}\}_{a'\in A_n}\right)=0$.

Therefore, we can bound the variance of $\sum_{a'\in A_n}\hat{Y}^t_{a,a'}$ with respect to Q^t_a as follows:

$$\mathrm{Var}_t\left[\sum_{a'\in A_n}\hat{Y}^t_{a,a'}\right]=\sum_{a'\in A_n}\mathrm{Var}_t\left[\hat{Y}^t_{a,a'}\right]\le\sum_{a'\in A_n}\mathbf{E}_{Q^t_a}\left[\left(\hat{Y}^t_{a,a'}\right)^2\right]=\sum_{a'\in A_n}\frac{q^t_{a,a'}(Y^t_{a,a'})^2}{(q^t_{a,a'}+\gamma)^2}\le\sum_{a'\in A_n}\frac{Y^t_{a,a'}}{q^t_{a,a'}+\gamma},$$

where the last inequality is due to $\frac{q^t_{a,a'}}{q^t_{a,a'}+\gamma}\le 1$ and $Y^t_{a,a'}\le 1$. Let $V_t:=\mathrm{Var}_t\left[\sum_{a'\in A_n}\hat{Y}^t_{a,a'}\right]$. Notice that $\mu_t+\gamma V_t\le\sum_{a'\in A_n}\frac{q^t_{a,a'}Y^t_{a,a'}+\gamma Y^t_{a,a'}}{q^t_{a,a'}+\gamma}=\sum_{a'\in A_n}Y^t_{a,a'}$. Then, we have the following inequality held:

$$\begin{aligned}&\sum_{a'\in A_n}\left(\hat{L}^T_{a,a'}-L^T_{a,a'}\right)=\sum_{a'\in A_n}\hat{L}^T_{a,a'}-\sum_{t=1}^T\sum_{a'\in A_n}Y^t_{a,a'}\\&\le\sum_{a'\in A_n}\hat{L}^T_{a,a'}-\sum_{t=1}^T(\mu_t+\gamma V_t)=\underbrace{\sum_{t=1}^T\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)}_{=:(\mathrm{c})}.\end{aligned}\tag{13}$$

Next, we bound Term (c) by using the Markov inequality (i.e., $P(X\ge\epsilon)\le\frac{\mathbf{E}[X]}{\epsilon}$ for any random variable $X>0$ and $\epsilon>0$) as follows:

$$\begin{aligned}&\Pr\left(\sum_{t=1}^T\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\ge\frac{1}{\gamma}\log\frac{1}{\delta}\right)=\Pr\left(\exp\left(\gamma\sum_{t=1}^T\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\right)\ge\frac{1}{\delta}\right)\\&\le\delta\mathbf{E}\left[\exp\left(\gamma\sum_{t=1}^T\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\right)\right]=\delta\mathbf{E}\left[\prod_{t=1}^T\exp\left(\gamma\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\right)\right]\\&=\delta\mathbf{E}\left[\prod_{t=1}^T\mathbf{E}_{Q^t_a}\left[\exp\left(\gamma\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\right)\right]\right],\end{aligned}\tag{14}$$

where the last equation is due to the law of total expectation and independence of meta-distributions. Then, using the facts that $\exp(x)\le 1+x+x^2$ for $x\le 1$ and $1+x\le\exp(x)$ for all $x\in\mathbb{R}$, we have

$$\begin{aligned}&\mathbf{E}_{Q^t_a}\left[\exp\left(\gamma\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)\right)\right]=\exp(-\gamma^2V_t)\mathbf{E}_{Q^t_a}\left[\exp\left(\gamma\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t\right)\right)\right]\\&\le\exp(-\gamma^2V_t)\left(1+\gamma\mathbf{E}_{Q^t_a}\left[\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t\right]+\gamma^2\mathbf{E}_{Q^t_a}\left[\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t\right)^2\right]\right)\\&=\exp(-\gamma^2V_t)\left(1+\gamma^2V_t\right)\le\exp(\gamma^2V_t-\gamma^2V_t)=1.\end{aligned}\tag{15}$$

Notice that $\left\{\sum_{s=1}^t\left[\sum_{a'\in A_n}\hat{Y}^s_{a,a'}-\mu_s\right]\right\}_{t\ge 0}$ is a martingale sequence. The above inequality actually gives a high-probability bound for the tail of the martingale sequence for T rounds in terms of the variance, i.e., $\sum_{t=1}^T V_t$. Then, by substituting (15) into (14), we have with probability at least $1-\delta$:

$$\sum_{t=1}^T\left(\sum_{a'\in A_n}\hat{Y}^t_{a,a'}-\mu_t-\gamma V_t\right)<\frac{1}{\gamma}\log\frac{1}{\delta},$$

and Lemma A.2 follows according to (13). □