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

Zhiming Huang Jianping Pan

Department of Computer Science, University of Victoria, Victoria BC, Canada

Abstract

- 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.
- 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.

Introduction

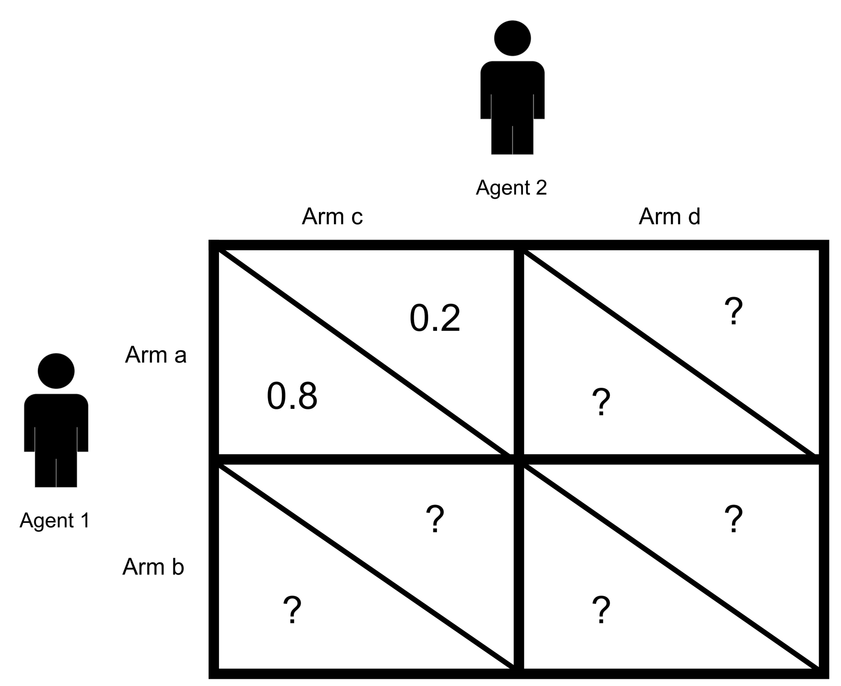


Figure 1: An example of MAB-UG with two agents and two arms for each agent.

We study the unknown general-sum games (i.e., black-box games) with bandit feedback repeated for T rounds, involving an agent set $\mathcal{N} := \{1, \dots, N\}$ 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. At each time $t = 1, \dots, T$:

- Each agent n plays an action $a_n^t \in A_n$
- Each agent n observes a reward $u_n(a_n^t; \mathbb{A}_{-n}^t)$, where
  - $u_n : \mathcal{A} \to [0, 1]$, mapping the actions of all agents to agent n's rewards
  - $(a_n^t; \mathbb{A}_{-n}^t)$ is an abbreviation of $\mathbb{A}^t := (a_1^t, \dots, a_n^t, \dots, a_N^t)$ with a highlight of agent n's action a_n against other agents' actions.
- The objective of each agent is to (1) accumulate as many rewards as possible and (2) achieve the ϵ-correlated equilibrium.

Challenges:

- Each agent does not know the underlying game structure nor the number of other agents.
- Each agent cannot observe the actions and rewards of other agents.

Applications:

- End-to-end congestion control in computer networks.
- Medium access control in wireless communications.

Swap Regret and Our Contributions

Introduced by [1], swap regret is a general regret definition comparing the learning algorithm with $K_n^{K_n}$ competitors:

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

where $F_n : A_n \to A_n$ takes $a \in A_n$ as input and outputs $a' \in A_n$, and $\mathcal{F}$ is a finite set of F_n. Minimizing swap regret can accumulate many rewards as possible and converge to the ϵ-correlated equilibrium.

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 [1]	$\Omega\left(\sqrt{TK_n}\right)$ [1]
$O\left(\sqrt{TK_n^2 \log(K_n)}\right)$, exp-time, pseudo-regret [2]	
$O\left(\sqrt{TK_n^2 \log(K_n)}\right)$, poly-time, pseudo-regret [3]	$\Omega\left(\sqrt{TK_n \log(K_n)}\right)$ [3]
$O\left(\sqrt{TK_n^2 \log(K_n/\delta)}\right)$, poly-time, conditionally expected regret [4]	
$O\left(\sqrt{TK_n^2 \log(K_n/\delta)}\right)$, 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, Corollary 5.4)	

The LCE-IX Algorithm

The LCE-IX Algorithm is based on the swap-regret-minimizing framework [1], calling the Exp3-IX algorithm [5] as subroutines. LCE-IX maintains K_n subroutines, and each K_n subroutine maintains a probability distribution $Q_a^t := \{q_{a,a'}^t : \forall a' \in A_n\}$ among K_n actions. Let $P_n^t := \left[p_1^t, \cdots, p_{K_n}^t\right]$ be the probability distribution of selecting an action $a_n \in A_n$, which is calculated by solving the following equations for P_n^t

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

The observed rewards are then distributed to subroutines according to their Q_a^t by $Y_{a,a'}^t := \frac{\mathbf{1}[a_n^t = a'] p_a^t q_{a,a'}^t}{p_{a'}^t}(1 - X_n^t)$, and estimated with the implicit exploration technique [5] by $\hat{Y}_{a,a'}^t := \frac{Y_{a,a'}^t}{q_{a,a'}^t + \gamma_t}$.

Then, Q_a^t is updated by following the Exp3 algorithm: $q_{a,a'}^{t+1} = \frac{\exp\left(-\eta_{t+1} \hat{L}_{a,a'}^t\right)}{\sum_{a'' \in A_n} \exp\left(-\eta_{t+1} \hat{L}_{a,a''}^t\right)}$.

Analytical Results

Theorem 5.3: Let $\delta \in (0, 1)$. With probability at least $1 - \delta$, $\eta_t = \sqrt{\frac{\log(K_n) + \log(K_n/\delta)}{t}}$ and $\gamma_t = \eta_t/2$, the instantaneous swap regret over T rounds is bounded by $O\left(\sqrt{TK_n^2 \log(K_n/\delta)}\right)$.

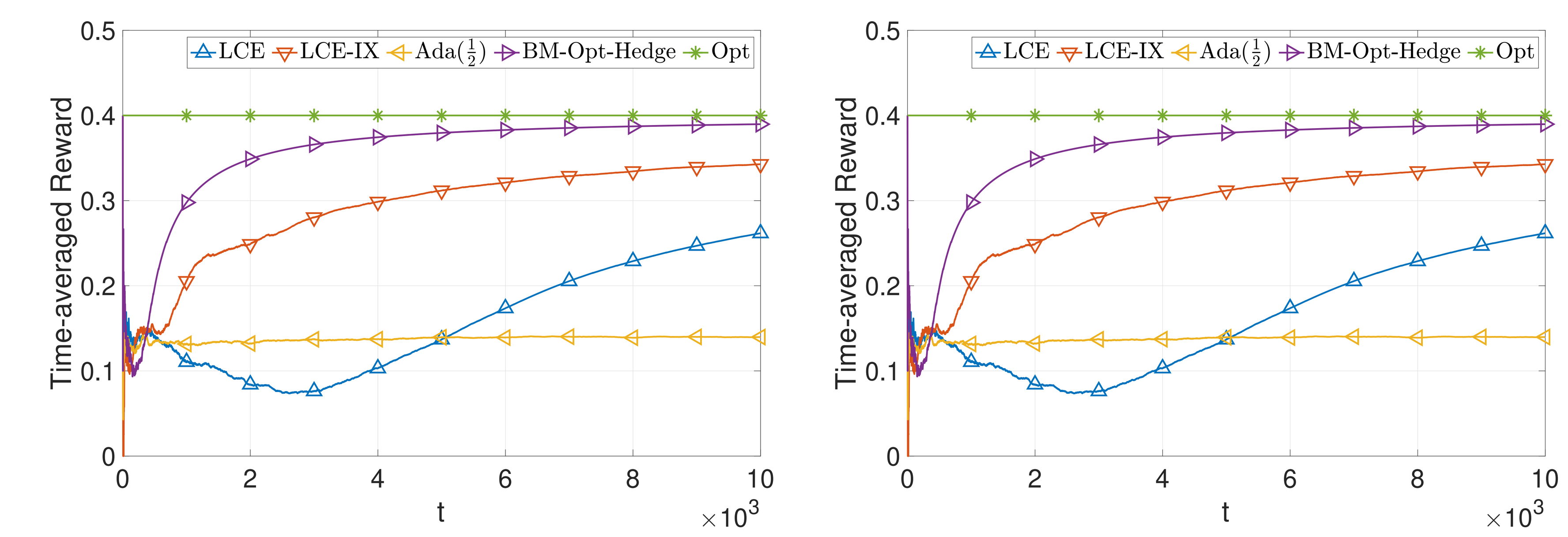
Theorem 5.4: With $\eta_t = \sqrt{\frac{\log(K_n)}{t}}$ and $\gamma_t = \eta_t/2$, the expected swap regret is bounded by $O\left(\sqrt{TK_n^2 \log(K_n)}\right)$.

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 $T \to \infty$, $\hat{\mathbf{P}}^T$ converges to the set of correlated equilibria almost surely.

Numerical Experiments

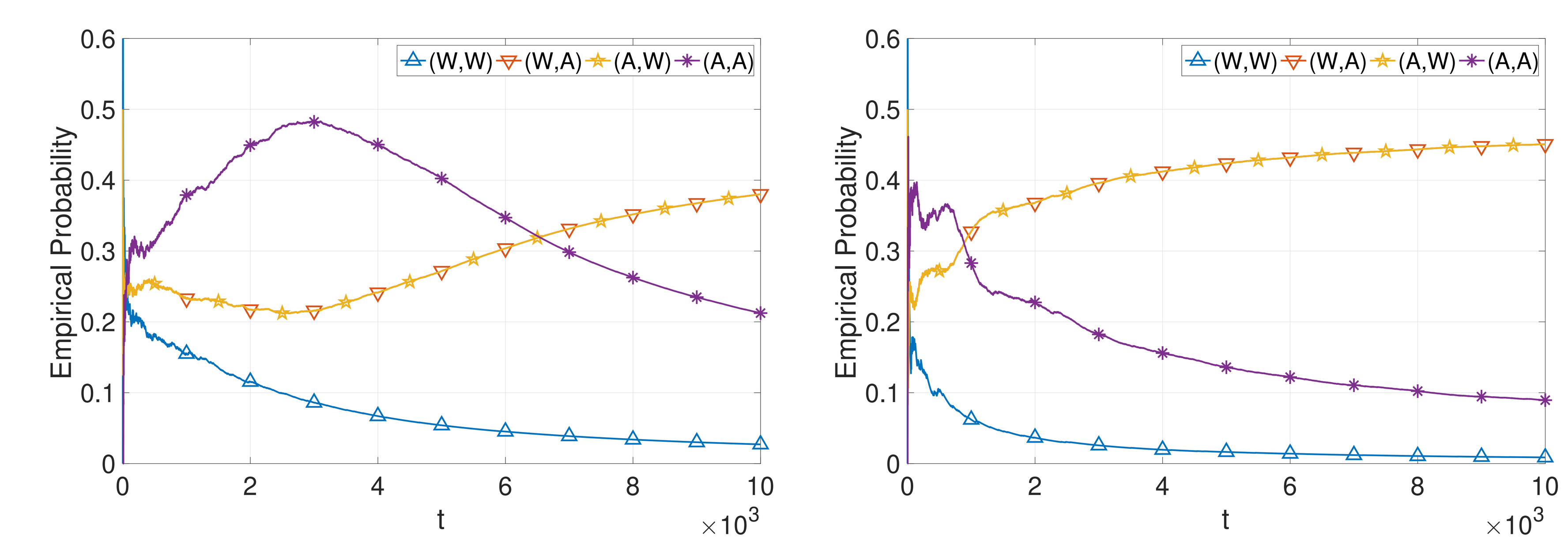
Table 2: The reward matrix for the medium access game

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



(a) Agent 1 (b) Agent 2

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



(a) LCE (b) LCE-IX

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

References

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

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

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

[4] C. Jin, Q. Liu, Y. Wang, and T. Yu, ``V-Learning--A Simple, Efficient, Decentralized Algorithm for Multiagent RL,'' in *Proc. ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

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