Individual Regret in Cooperative Nonstochastic Multi-Armed Bandits

Yogev Bar-On and Yishay Mansour

NeuRIPS 2019

Tel Aviv University

Table of Contents

Introduction

The Nonstochastic Multi-Armed Bandit Setting

- The problem is played by N agents over a finite action set $A = \{1, \ldots, K\}$.
- The agents $V = \{1, \ldots, N\}$ are communicating over a communication graph $G = \langle V, E \rangle$.
- At each time step $t = 1, 2, \ldots, T$, each agent $v \in V$ picks an action $I_t(v) \in A$. It then suffers some loss $\ell_t(I_t(v)) \in [0, 1]$ chosen by an adversary in advance.
- Agents observe only the loss of their chosen action. However, agents may receive information from their neighbors, and thus have access to other losses as well.

Regret Minimization

Each agent's goal is to minimize its *expected regret* over T steps, compared to the best action in hindsight:

$$R_T(v) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t(v)) - \min_{i \in A} \sum_{t=1}^{T} \ell_t(i)\right].$$

- Previous work bounded the **average** expected regret $\frac{1}{N}\sum_{v \in V} R_T(v)$.
- In this work, we bound the regret of each agent **individually**, and the bound holds for all agents simultaneously.

The Exp3-Coop Algorithm

Cesa-Bianchi et al. [2019][1] generalized Exp3 to the cooperative setting with an algorithm called Exp3-Coop.

- Agents share with their neighbors in the communication graph each step their loss and their probability vector.
- Exp3-Coop provably achieves an **average** expected regret of:

$$\widetilde{O}\left(\sqrt{\left(1+\frac{K}{N}\alpha\left(G\right)\right)T}\right),$$

where $\alpha\left(G\right)$ is the independence number of G.

[1] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. The Journal of Machine Learning Research, 20(1):613–650, 2019.

The Center-Based Policy

The Center-Based Policy

To obtain a low **individual** regret, we present the center-based policy.

- The communication graph is artificially partitioned to connected components (with methods that will be presented later).
- There is special set of *center agents* C, such that each component contains a single center.
- Neighbors of center agents (called *center-adjacent*) must be in the component of their neighboring center.
- Centers use Exp3-Coop, and other agents copy (with delay, through other agents in the component) their weights vector from their center - the center in their component.

Mass

Definition (Mass)

The *mass* of a center agent $c \in C$ is defined to be

$$M(c) \equiv \min\{\deg(c) + 1, K\},$$

and the mass of non-center agent $v \in V \setminus C$ is

$$M(v) \equiv e^{-\frac{1}{6}d(v)} M(\mathcal{C}(v)).$$

Theorem (Center-Based Regret)

If all agents use the center-based policy, the regret of any agent $v \in V$ holds:

$$R_T(v) = O\left(\sqrt{(\ln K)\frac{K}{M(v)}T}\right).$$

Partitioning the Communication Graph

Partitioning Settings

Our objective is to partition the communication graph to components such that each agent will have a large mass (in the order of its degree). There are two settings in which we can do this:

- **The informed setting** - Each agent has a complete knowledge of the graph structure, and the partitioning can be computed ahead of time.
- **The uninformed setting** - Each agent only has information about its neighborhood, and an upper bound on the total number of agents. In this setting, the partitioning must occur after the game has started, and thus there is a penalty to the regret.

Partitioning Algorithms

The following algorithms are used to partition the communication graph:

1. First, we choose which agents will be centers.
   - *Compute-Centers-Informed* is used in the informed setting.
   - *Compute-Centers-Uninformed* is used in the uninformed setting.
2. Then, we assign a component for each non-center agent, keeping the requirements for the center-based policy.
   - *Centers-to-Components* is used in both the informed and uninformed settings.

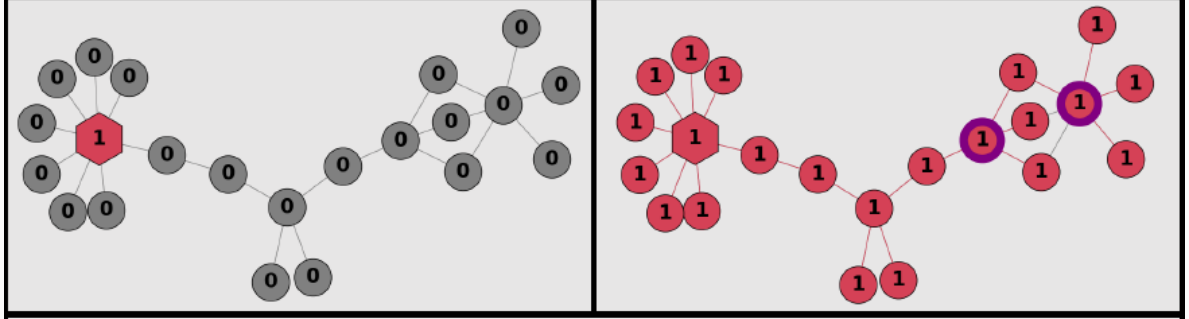
Theorem (Main Result)

Using our partitioning algorithms and the center-based policy, the individual expected regret of all agents $v \in V$ holds:

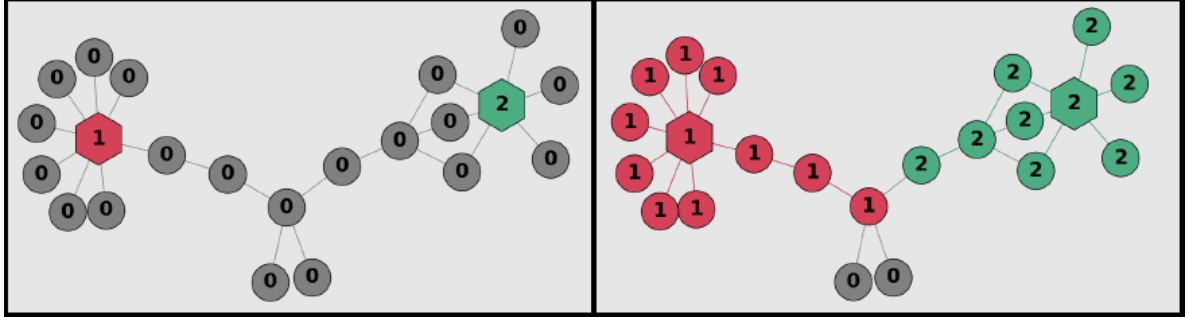
$$R_T(v) = \widetilde{O}\left(\sqrt{(\ln K)\left(1 + \frac{K}{\deg(v)}\right)T}\right).$$

Illustration

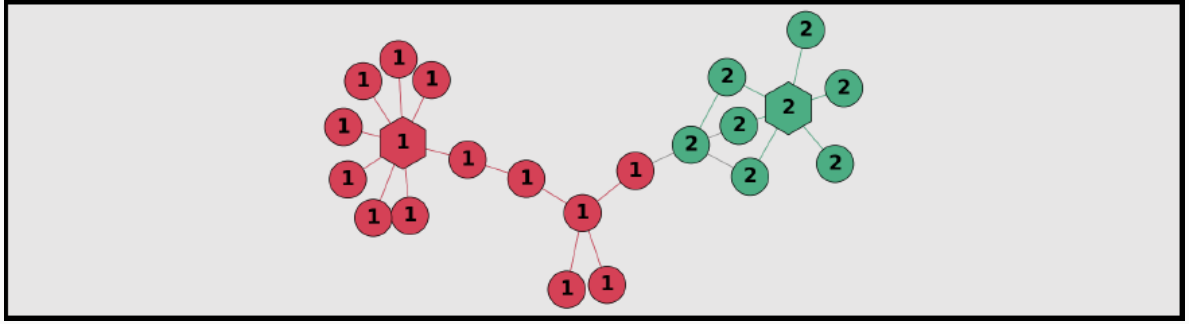
First, the agent with the highest degree becomes a center.
Then, every other agent must be in its component.



Now, the unsatisfied agent (marked with a large purple border) with the highest degree becomes center.
Then, non-center agents start choosing components, first of their closest center.



Finally, agents may switch to a farther center if doing so will lower their expected regret:



Conclusion

Conclusion

- We investigated the cooperative nonstochastic multi-armed bandit problem, and presented the center-based cooperation policy.
- We provided partitioning algorithms that provably yield a low individual regret bound that holds simultaneously for all agents.
- We express this bound in terms of the agents' degree in the communication graph. This bound strictly improves a previous regret bound from [Cesa-Bianchi et al., 2019], and also resolves an open question from that paper.

Questions?