Decentralized Multi-Armed Bandit with Imperfect Observations

Keqin Liu*, Qing Zhao*, Bhaskar Krishnamachari◊

*University of California, Davis
◊University of Southern California, Los Angeles

Supported by ARO and NSF.
Classic MAB with A Single Player
The Problem

\(N\)-Armed Bandit

- \(N\) arms and a single player
- select one arm to play at each time
- Each arm, when played, yields reward with unknown mean.
- Maximize the long-run reward.

Exploitation V.S. Exploration

- Exploitation: play an arm appearing to be the best for immediate reward
- Exploration: play an arm for learning to minimize future mistakes.
Regret

Performance Measure: Regret

- $\Theta \overset{\Delta}{=} (\theta_1, \ldots, \theta_N)$: unknown parameter set
- $V_T^\pi(\Theta)$: total reward of policy $\pi$
- $T\theta_{max}$: max total reward if $\Theta$ is known
- Regret (cost of learning):

$$R_T^\pi(\Theta) \overset{\Delta}{=} T\theta_{max} - V_T^\pi(\Theta) = \sum_{i: \theta_i < \theta_{max}} (\theta_{max} - \theta_i)E[\text{time spent on } \theta_i]$$

Objective: minimize the rate that $R_T^\pi(\Theta)$ grows as $T \to \infty$.

Sub-linear regret$\implies$Maximum Average Reward
Minimum Regret Growth Rate

- **Lai&Robbins’85:**
  \[
  R_T^*(\Theta) \sim \sum_{i: \theta_i < \theta_{max}} \theta_{max} - \theta_i \frac{I(\theta_i, \theta_{max})}{\log T} \text{ as } T \to \infty
  \]

  - An optimal policy constructed to achieve the minimum regret rate.

- **Anantharam&Varaiya&Walrand’87:**
  - Extension from single play to multiple plays
  - Extension to Markovian reward process.

- **Agrawal’95, Auer&Cesa-Bianchi&Fischer&Informatik’02:**
  - Order-Optimal Index Policies.
Decentralized MAB with Multiple Players
Problem Formulation

- $M \ (M < N)$ distributed players

- Each player selects one arm to play

- Each Player makes decisions based on its local observations without information exchange

- Colliding players either share the reward or receive no reward.

Exploitation V.S. Exploration V.S. Competition
System Regret

- $V^\pi_T(\Theta)$: the total reward under a decentralized policy $\pi$.

- $T\sum_{i=1}^{M}\theta_{\sigma(i)}$: max total reward with known $\Theta$ and centralized scheduling.

- The regret $R^\pi_T(\Theta) \overset{\Delta}{=} T\sum_{i=1}^{M}\theta_{\sigma(i)} - V^\pi_T(\Theta)$.

The regret is defined w.r.t. the same upper bound as the centralized case.
Optimal Regret of Decentralized MAB

Optimal Regret (Liu&Zhao’09)

- The minimum regret rate is logarithmic
  \[ R_T^*(\Theta) \sim C(\Theta) \log T \quad \text{as } T \to \infty \]

- A framework of constructing order-optimal and fair decentralized policies was proposed under general reward, observation, and collision models.

Independent Work (Anandkumar&Michael&Tang)

- Focus on Bernoulli distribution

- Fairness only in a probabilistic sense.
Imperfect Observations

Each player can only observe the reward received

- A collision affects not only the reward received, but also the information obtained for learning.
Cognitive Radio for Dynamic Spectrum Access

- $N$ channels and $M$ distributed secondary users.
- Channel state: Bernoulli (busy 0 or idle 1) with unknown mean
- **Imperfect Sensing:** an idle channel can be sensed as busy, *vice versa*
- Primary collision and secondary collision are indistinguishable.
Multi-Agent Systems

- $M$ distributed agents search/collect targets in $N$ locations

- Each agent only observes the reward received

- Reward received in a collision does not reflect the quality of the location.
Implications of Imperfect Observation

- Collisions are not observable

- Including corrupted observations in learning can lead to misidentified arm rank
Time is partitioned into **Exploration** and **Exploitation** subsequences.

- **Exploration**: Players play all arms in a round-robin fashion with different offsets
- **Exploitation**: Each player plays the $M$ best arms learned so far
The Optimal Denseness of Exploration

- Time is partitioned into Exploration and Exploitation subsequences.

- **Exploration**: Players play all arms in a round-robin fashion with different offsets

- **Exploitation**: Each player plays the $M$ best arms learned so far

- Optimal denseness of exploration
  - Too dense: spending too much time on bad arms
  - Too sparse: not enough observations for learning
The Optimal Denseness of Exploration

Logarithmic?
The Optimal Denseness of Exploration

▶ Logarithmic?

▶ How about $O(\sqrt{T})$?
By choosing the exploration denseness $O(\sqrt{T})$ with total time $T$, the system regret achieves the same $O(\sqrt{T})$ order

- **Fairness**: each player achieves the same local average reward at the same rate

- Achieving the maximum average reward
Cognitive Radio for Dynamic Spectrum Access

Achieving Logarithmic Regret (Liu&Zhao&Krishnamachari’10)

- The sensing outcome at each Tx is not affected by collisions
- Control information exchange between each Tx and Rx to ensure synchronized learning.
Conclusion

- Formulated decentralized MAB with imperfect observations
- Design the denseness of the exploration sequence to balance tradeoff between exploration and exploitation
- Achieve $O(\sqrt{T})$ regret with fairness
- Achieve logarithmic regret for a special case (cognitive radio)
Conclusion

- Formulated decentralized MAB with imperfect observations
- Design the denseness of the exploration sequence to balance tradeoff between exploration and exploitation
- Achieve $O(\sqrt{T})$ regret with fairness
- Achieve logarithmic regret for a special case (cognitive radio)

Extension (H. Liu & K. Liu & Q. Zhao’10)

- Achieve logarithmic regret for Markovian rewards (perfect observation)
Application Example—Ore Mining

- $M$ distributed agents assigned to ore mining among $N$ locations

- The reward distribution considered as log-Gaussian distributed with unknown mean and variance (Sharp’76)

- If multi agents choose the same location, they share the reward in an unknown way (e.g., depending on which agent is faster in collection)
Application Example—Ore Mining

Exploration subsequence: cardinality $\sim a(\sqrt{T})$