Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms

Xuchuang Wang[1], Hong Xie[2], John C.S. Lui[1]

The Chinese University of Hong Kong[1], Chongqing University[2]

June 27, 2022

Multiple-Play Multi-Armed Bandits

- K arms: each associated with a reward random variable X_k with mean μ_k.
  - Assume $\mu_1 > \cdots > \mu_N > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls N arms among $\in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$ from N pulled arms.
- Denote action $\boldsymbol{a}_t \in \mathbb{N}_+^K$: if arm k is pulled then $a_{k,t} = 1$; or otherwise $a_{k,t} = 0$.
  - e.g., $\boldsymbol{a}_t = (0, 1, 1, 0, \ldots)$
  - $\sum_{k=1}^K a_{k,t} = N$
- Goal: maximize total reward; or minimize the regret

$$\mathbb{E}[\text{Reg}(T)] := \underbrace{\qquad}_{\text{Optimal}} - \underbrace{\qquad}_{\text{Algorithm's}}.$$

Multiple-Play Multi-Armed Bandits

- K arms: each associated with a reward random variable X_k with mean μ_k.
  - Assume $\mu_1 > \cdots > \mu_N > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls N arms among $\in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$ from N pulled arms.
- Denote action $\boldsymbol{a}_t \in \mathbb{N}_+^K$: if arm k is pulled then $a_{k,t} = 1$; or otherwise $a_{k,t} = 0$.
  - e.g., $\boldsymbol{a}_t = (0, 1, 1, 0, \ldots)$
  - $\sum_{k=1}^K a_{k,t} = N$
- Goal: maximize total reward; or minimize the regret

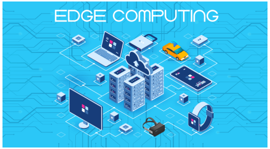
$$\mathbb{E}[\mathrm{Reg}(T)] := \underbrace{T \sum_{k=1}^{N} \mu_k}_{\text{Optimal}} - \underbrace{}_{\text{Algorithm's}}.$$

Multiple-Play Multi-Armed Bandits

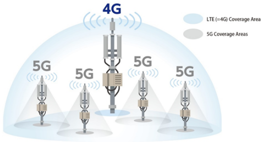
- K arms: each associated with a reward random variable X_k with mean μ_k.
  - Assume $\mu_1 > \cdots > \mu_N > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls N arms among $\in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$ from N pulled arms.
- Denote action $\boldsymbol{a}_t \in \mathbb{N}_+^K$: if arm k is pulled then $a_{k,t} = 1$; or otherwise $a_{k,t} = 0$.
  - e.g., $\boldsymbol{a}_t = (0, 1, 1, 0, \ldots)$
  - $\sum_{k=1}^K a_{k,t} = N$
- Goal: maximize total reward; or minimize the regret

$$\mathbb{E}[\mathrm{Reg}(T)] := \underbrace{T \sum_{k=1}^{N} \mu_k}_{\text{Optimal}} - \underbrace{\sum_{t=1}^{T} \sum_{k: a_{k,t}=1} \mu_k}_{\text{Algorithm's}}.$$

Shareable Finte-Capacity Arm



(a) Edge Computing [2]



(b) Wireless Network [1]

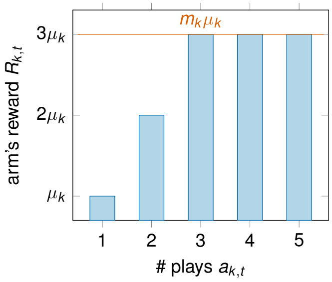
- Each arm has **two unknowns**:
  - "per-load" reward mean μ_k and integer reward capacity m_k.

Shareable Finte-Capacity Arm

- Each arm has **two unknowns**:
  - "per-load" reward mean μ_k and integer reward capacity m_k.
- If $a_{k,t}$ plays pull the arm k with m_k capcity, then the reward from this arm

$$R_{k,t} := \min\{a_{k,t}, m_k\} X_{k,t} = \begin{cases} a_{k,t} X_{k,t}, & a_{k,t} \leqslant m_k \\ m_k X_{k,t}, & a_{k,t} > m_k \end{cases}$$

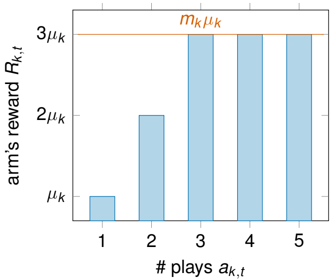
  - $X_{k,t}$ is the "per-load" reward random variable.



Shareable Finte-Capacity Arm

- Each arm has **two unknowns**:
  - “per-load” reward mean μ_k and integer reward capacity m_k.
- If $a_{k,t}$ plays pull the arm k with m_k capcity, then the reward from this arm

$$R_{k,t} := \min\{a_{k,t}, m_k\}X_{k,t} = \begin{cases} a_{k,t}X_{k,t}, & a_{k,t} \leqslant m_k \\ m_k X_{k,t}, & a_{k,t} > m_k \end{cases}$$



  - $X_{k,t}$ is the “per-load” reward random variable.
- Optimal allocation:

$$\boldsymbol{a}^* := \left(m_1, \dots, m_{L-1}, N - \sum_{k=1}^{L-1} m_k, 0, \dots, 0\right)$$

where $L := \min\left\{n : \sum_{k=1}^{n} m_k \geqslant N\right\}$., the smallest number of top arms covering N plays.

Learn Reward Capacity m_k

- Sample Complexity Minimax Lower Bound (Gaussian): for any estimator $\hat{m}_t$

$$n \geqslant \frac{\sigma_k^2 m_k^2 \log(1/4\delta)}{\mu_k^2}.$$

Explorations can have any number of plays pulling the same arm.

Learn Reward Capacity m_k

- Sample Complexity Minimax Lower Bound (Gaussian): for any estimator $\hat{m}_t$

$$n \geqslant \frac{\sigma_k^2 m_k^2 \log(1/4\delta)}{\mu_k^2}.$$

  Explorations can have any number of plays pulling the same arm.

- Estimator: $\hat{m}_t = \frac{\text{“full-load” } \hat{\nu}_{k,t}}{\text{“per-load” } \hat{\mu}_{k,t}} \left(\approx \frac{m_k \mu_k}{\mu_k}\right)$
  - Individual exploration (IE, $a_{k,t} < m_k$) $\implies$ “per-load” reward empirical mean $\hat{\mu}_{k,t}$
  - United exploration (UE, $a_{k,t} \geqslant m_k$) $\implies$ “full-load” reward empirical mean $\hat{\nu}_{k,t}$

Learn Reward Capacity m_k

- Sample Complexity Minimax Lower Bound (Gaussian): for any estimator $\hat{m}_t$

$$n \geqslant \frac{\sigma_k^2 m_k^2 \log(1/4\delta)}{\mu_k^2}.$$

Explorations can have any number of plays pulling the same arm.

- Estimator: $\hat{m}_t = \frac{\text{“full-load” } \hat{\nu}_{k,t}}{\text{“per-load” } \hat{\mu}_{k,t}} \left(\approx \frac{m_k \mu_k}{\mu_k} \right)$
  - Individual exploration (IE, $a_{k,t} < m_k$) $\implies$ “per-load” reward empirical mean $\hat{\mu}_{k,t}$
  - United exploration (UE, $a_{k,t} \geqslant m_k$) $\implies$ “full-load” reward empirical mean $\hat{\nu}_{k,t}$
- Estimator's Sample Complexity Upper Bound: $\tau_{k,t}$ IEs and $\iota_{k,t}$ UEs

$$\tau_{k,t}, \iota_{k,t} \leqslant \frac{49 m_k^2 \log(2/\delta)}{\mu_k^2}.$$

Regret Minimization for MP-MAB with Shareable Arms

- Regret Lower Bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geqslant \underbrace{\sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)}}_{\text{estimate reward mean}} + \underbrace{\sum_{k=1}^{L-1} \frac{\Delta_{k,L}\sigma^2 m_k^2}{\mu_k^2} + \frac{\Delta_{L,L+1}\sigma^2 m_L^2}{(m_L - \bar{m}_L + 1)^2 \mu_L^2}}_{\text{estimate reward capacity}}$$

Regret Minimization for MP-MAB with Shareable Arms

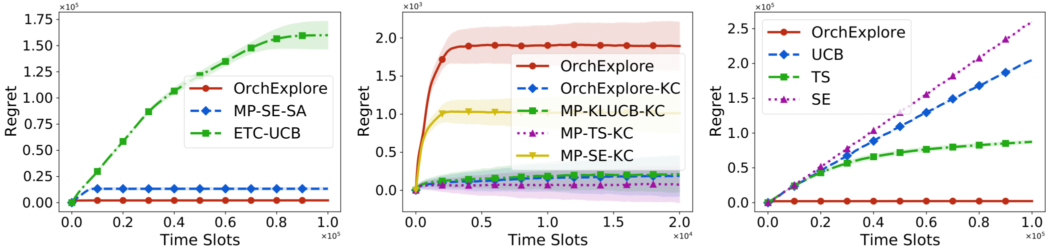
- Regret Lower Bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geqslant \underbrace{\sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)}}_{\text{estimate reward mean}} + \underbrace{\sum_{k=1}^{L-1} \frac{\Delta_{k,L}\sigma^2 m_k^2}{\mu_k^2} + \frac{\Delta_{L,L+1}\sigma^2 m_L^2}{(m_L - \bar{m}_L + 1)^2 \mu_L^2}}_{\text{estimate reward capacity}}$$

- `OrchExplore` Algorithm: Parsimonious IEs + UEs
- Regret Upper Bound

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \leqslant \sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)} + \sum_{k=1}^{L-1} \frac{49 w_k m_k^2}{\mu_k^2} + \frac{49 w_L m_L^2}{(m_L - \bar{m}_L + 1)^2 \mu_L^2}.$$

Simulations



(a) `OrchExplore` *vs.* Others (b) Price of learning m_k (c) Implicitly learn m_k

Thank you!

Full paper at arXiv:2206.08776

References I

[1] Tokyu Corporation and Sumitomo Corporation. Launch of pilot experiment on 5g base-station-sharing business in shibuya, 2019. URL https://www.sumitomocorp.com/en/africa/news/release/2019/group/12330.

[2] SPEC INDIA. What is edge computing? the quick overview explained with examples, 2019. URL https://www.spec-india.com/blog/what-is-edge-computing-the-quick-overview-explained-with-examp