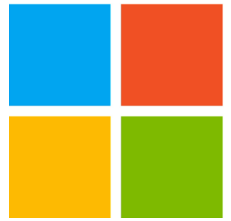
上海交通大学
约翰·霍普克罗夫特
计算机科学中心
John Hopcroft Center for Computer Science



Simultaneously Learning Stochastic and Adversarial Bandits with General Graph Feedback (ICML 2022)

Fang Kong, Yichi Zhou, Shuai Li

What are bandits?

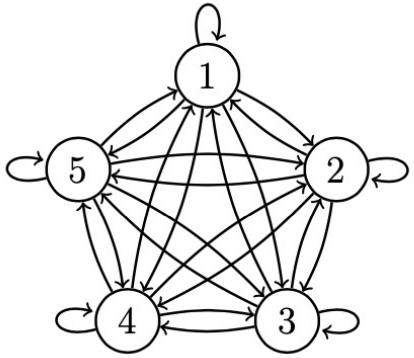
Time	1	2	3	4	5	6	7	8	9	10	11	12
Left arm	$1	$0			$1	$1	$0					
Right arm			$1	$0								

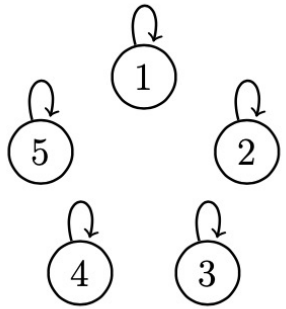
To accumulate as many rewards, which arm would you choose next?

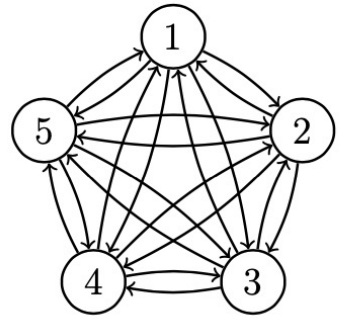
Exploitation V.S. Exploration

- Lattimore, Tor, and Csaba Szepesvári. Bandit Algorithms. Cambridge University Press, 2020

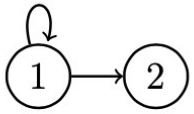
Interactive framework

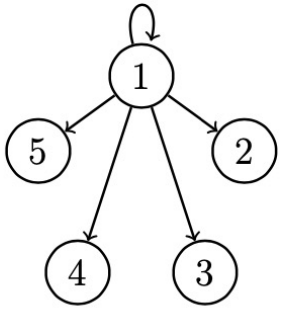


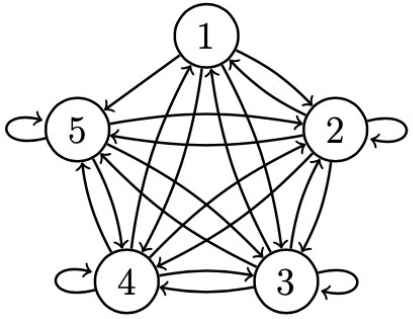




- For $t = 1,2, \ldots$
  - The agent selects arm $I_t \in V$
  - The environment produces reward $r_t = (r_t(1), r_t(2), \ldots, r_t(K)) \in [0,1]^K$
  - The agent observes $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - The agent receives reward $r_t(I_t)$







- Feedback graph $G = (V, E)$
  - Arm set $V = \{1,2, \ldots, K\}$
  - Edge set $E = \{(i, j)\}$

- Alon, Noga, et al. “Online learning with feedback graphs: Beyond bandits.” Conference on Learning Theory. PMLR, 2015.

Objectives for two environments

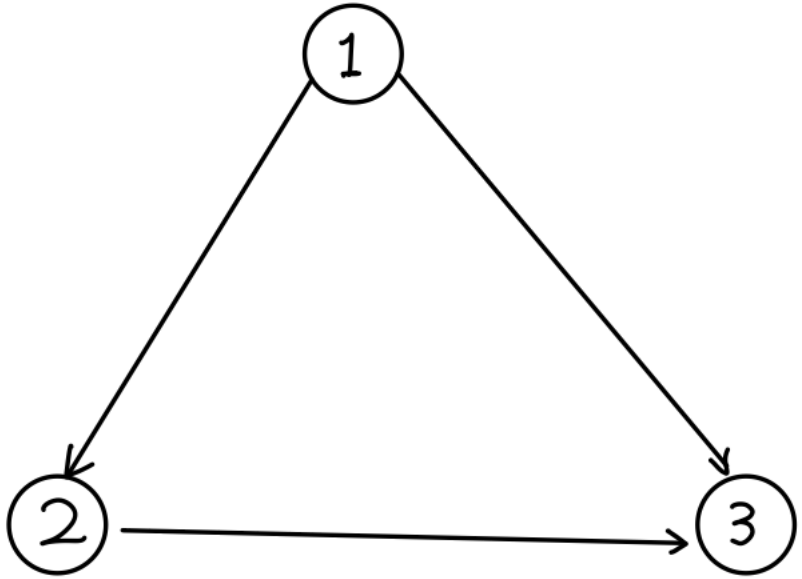
- Stochastic setting
  - $r_t(i)$ is drawn independently from a fixed distribution
  - $\mathbb{E}[r_t(i)\,] = \mu_i$
  - Aim to minimize the regret

$$Reg(T) = \max_{i \in V} \sum_{t=1}^{T} (\mu_i - \mu_{I_t}) := \sum_{t=1}^{T} (\mu_{i^*} - \mu_{I_t}) := \sum_{t=1}^{T} \Delta_{I_t}$$

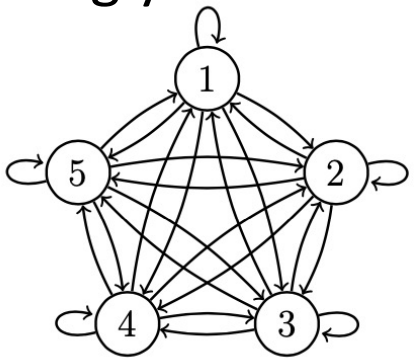
- Adversarial setting
  - $r_t(i)$ can be chosen arbitrarily by an adversary

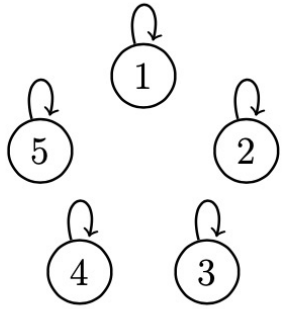
$$Reg(T) = \max_{i \in V} \sum_{t=1}^{T} r_t(i) - \sum_{t=1}^{T} r_t(I_t)$$

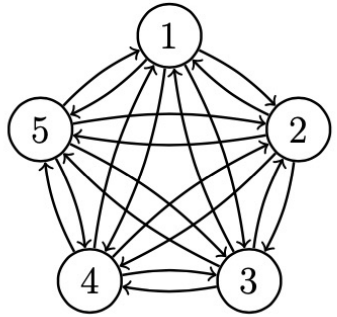
Observability



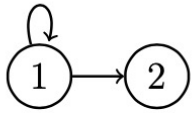
- Strongly observable:

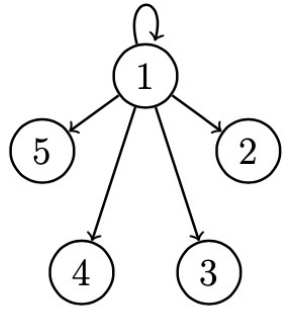


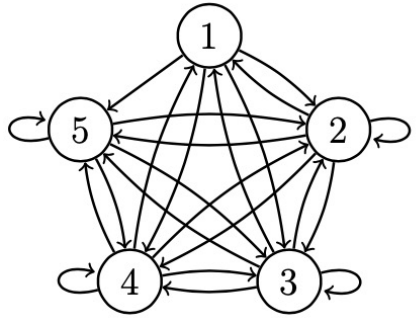




- Weakly observable:







- The agent cannot determine which arm is optimal will suffer $O(T)$ regret
- We consider observable graphs, i.e., $N^{in}(i) \neq \emptyset, \forall i$

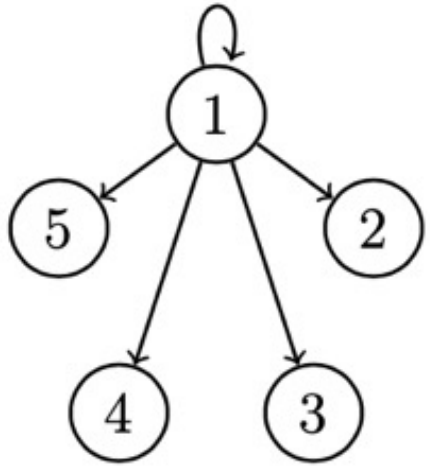
Previous results

	Stochastic	Adversarial
Wu et al. (2015)	$\Theta\left(\frac{\log T}{\Delta^2}\right)$	
Alon et al. (2015)		$\Theta\left(T^{2/3}\right)$
Chen et al. (2021)		$\Theta\left(T^{2/3}\right)$

- Can we achieve best-of-both-worlds guarantees?
  - Erez et al. (2021) also try to solve this problem but only for undirected graph with self-loops.

A simple idea of explore-then-commit (ETC)

- In the stochastic setting
  - Sample dominating set (arm 1) until all sub-optimal arms are identified and then focus on the optimal one
  - Each sub-optimal arm i needs to be observed for $O(\log T/\Delta_i^2)$ times
  - $O(|D|\log T/\Delta^2)$



- However, it would fail in the adversarial setting: $O(T)$
  - The optimal arm changes with the horizon

“Switch” algorithm

Only exploration would fail for rounds before switch in adversarial setting

- For t=1,2,3,…
  - Determine $p_t(i)$ for each arm $i \in V$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - Detect whether the environment is adversarial
    - If true, run Exp3.G (Alon et al. 2015)

Add exploitation to bound in adversarial

- For t=1,2,3,…
  - For each arm $i \in V$
    - $p_{t,D}(i) = \frac{1}{|D|}\mathbb{I}\{i \in D\}$
    - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$
    - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma) p_{t,A}(i)$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - Detect whether the environment is adversarial
    - If true, run Exp3.G (Alon et al. 2015)

Optimize in the stochastic setting

- For t=1,2,3,…
  - For each arm $i \in V$
    - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$
    - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}$
    - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma)p_{t,A}(i)$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - Detect whether an arm in A is sub-optimal
    - If true, delete this arm from A
  - Detect whether the environment is adv:
    - If true, run Exp3.G (Alon et al. 2015)

Guarantee observations to detect adversarial

- For t=1,2,3,…
  - For each arm $i \in V$
    - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$
    - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j \in D \setminus D_A} p_D^{\text{old}}(j)\frac{\tau_j^D}{t}\right) + p_D^{\text{old}}(i)\frac{\tau_i^D}{t}\mathbb{I}\{i \in D \setminus D_A\}$
    - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma)p_{t,A}(i)$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - Detect whether an arm in A is sub-optimal
    - If true, delete this arm from A
  - Detect whether the environment is adversarial
    - If true, run Exp3.G (Alon et al. 2015)

Concentrations for detection

- Construct unbiased estimator for $r_t(i)$
  - $\tilde{r}_t(i) = r_t(i)\frac{\mathbb{I}\{i \in N^{out}(I_t)\}}{\sum_{j \in N^{in}(i)} p_t(j)}$
  - The averaged estimated reward for i at t is
  - $\widetilde{H}_t(i) = \frac{1}{t}\sum_{s=1}^{t} \tilde{r}_s(i)$
  - $\left|\widetilde{H}_t(i) - \mu_i\right| \leq \text{radius}_t(i) = O\left(\sqrt{\frac{1}{t\gamma_t}}\right)$ in stochastic setting
  - $\left|\widetilde{H}_t(i) - \frac{1}{t}\sum_{s=1}^{t} r_s(i)\right| \leq \text{radius}_t(i) = O\left(\sqrt{\frac{1}{t\gamma_t}}\right)$ in adversarial setting

Detect sub-optimal arms

- For t=1,2,3,…
  - For each arm $i \in V$
    - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$
    - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j \in D \setminus D_A} p_D^{\text{old}}(j)\frac{\tau_j^D}{t}\right) + p_D^{\text{old}}(i)\frac{\tau_i^D}{t}\mathbb{I}\{i \in D \setminus D_A\}$
    - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma)p_{t,A}(i)$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - If $\exists i$ such that $\widetilde{H}_t(j) - \text{radius}_t(j) > \widetilde{H}_t(i) + \text{radius}_t(i)$ where $j \in \text{argmax}_{j' \in A}\widetilde{H}_t(j')$
    - Delete arm i from A
  - Detect whether the environment is adversarial
    - If true, run Exp3.G (Alon et al. 2015)

Detect adversarial

- For t=1,2,3,…
  - For each arm $i \in V$
    - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$
    - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j \in D \setminus D_A} p_D^{\text{old}}(j)\frac{\tau_j^D}{t}\right) + p_D^{\text{old}}(i)\frac{\tau_i^D}{t}\,\mathbb{I}\{i \in D \setminus D_A\}$
    - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma)p_{t,A}(i)$
  - Sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - If $\exists i$ such that $\widetilde{H}_t(j) - \text{radius}_t(j) > \widetilde{H}_t(i) + \text{radius}_t(i)$ where $j \in \text{argmax}_{j' \in A}\widetilde{H}_t(j')$
    - Delete arm i from A
  - If $\exists i \notin A$ such that $\widetilde{H}_t(j) - \text{radius}_t(j) < \widetilde{H}_t(i) + \text{radius}_t(i)$ where $j \in \text{argmax}_{j' \in A}\widetilde{H}_t(j')$
    - A previous deleted arm becomes better-> adversarial
    - If true, run Exp3.G (Alon et al. 2015)

Regret analysis in adversarial setting

$$Reg(T) = \max_{i \in V} \sum_{t=1}^{T} (r_t(i) - r_t(I_t))$$
$$\leq \max_{i \in V} \sum_{t=1}^{\tau} (r_t(i) - r_t(I_t)) + \max_{i \in V} \sum_{t=\tau+1}^{T} (r_t(i) - r_t(I_t))$$

- During first τ rounds: Let $i^* \in \operatorname{argmax}_i \sum_{t=1}^{\tau} r_t(i)$
  - $i^* \in A_\tau$
  - $H_\tau(i^*) - H_\tau(i) < \widetilde{H}_\tau(i^*) - \widetilde{H}_\tau(i) + \text{radius}_\tau(i^*) + \text{radius}_\tau(i) < O(\text{radius}_\tau(i))$
  - Choosing $\gamma = O(t^{-1/3})$ to get regret $O(\tau^{2/3})$

- For the following rounds: $O(T^{2/3})$

Regret analysis in stochastic setting

- Detection for adversarial setting never satisfies

$$Reg(T) = \sum_{i \in V} \Delta_i \, N_i(T)$$

$$\leq \sum_{i \in V} \Delta_i \left(\tau_i + \sum_{t=1}^{\tau_i^D} \gamma_t + \sum_{t=\tau_i^D}^{T} \gamma_t \frac{\tau_i^D}{t} \right)$$

- $p_{t,A}(i) = \frac{1}{|A|} \mathbb{I}\{i \in A\}$;
- $p_{t,D}(i) = \frac{1}{|D_A|} \mathbb{I}\{i \in D_A\} \left(1 - \sum_{j \in D \setminus D_A} p_D^{\text{old}}(j) \frac{\tau_j^D}{t}\right) + p_D^{\text{old}}(i) \frac{\tau_i^D}{t} \mathbb{I}\{i \in D \setminus D_A\}$
- $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma) p_{t,A}(i)$

- $\tau_i = O\left(\frac{\log T}{\Delta_i^2}\right)^{3/2}$; $\tau_i^D = \max_{j \in N^{out}(i)} \tau_j$;
- $\sum_{t=1}^{\tau_i^D} \gamma_t = (\tau_i^D)^{2/3} = \max_{j \in N^{out}(i)} \log T \, / \Delta_j^2$
- $\tau_i^D \sum_{t=\tau_i^D}^{T} \frac{\gamma_t}{t} = O(\tau_i^D) = \tilde{O}\left(\frac{\log T}{\Delta_i^2}\right)^{3/2}$

Conclusion (general feedback graph)

	Stochastic	Adversarial
Wu et al. (2015)	$O(\lvert D\rvert \log T\,/\Delta^2)$	
Alon et al. (2015)		$O\left((\lvert D\rvert \log K)^{1/3}T^{2/3}\right)$
Chen et al. (2021)		$O\left((\delta \log K)^{1/3}T^{2/3}\right)$
Ours	$O\left(\lvert D\rvert^2(\log T/\Delta^2)^{3/2}\right)$	$O\left((\lvert D\rvert K^2)^{1/3}T^{2/3}\sqrt{\log T}\right)$

Future work

	Stochastic	Adversarial
Wu et al. (2015)	$O(\lvert D\rvert \log T\ /\Delta^2)$	
Alon et al. (2015)		$O\left((\lvert D\rvert \log K)^{1/3}T^{2/3}\right)$
Chen et al. (2021)		$O\left((\delta \log K)^{1/3}T^{2/3}\right)$
Ours	$O\left(\lvert D\rvert^2(\log T/\Delta^2)^{3/2}\right)$	$O\left((\lvert D\rvert K^2)^{1/3}T^{2/3}\sqrt{\log T}\right)$
Ito et al. (2022)	$O(\lvert D\rvert \log^2 T/\Delta^2)$	$O\left(D^{1/3}T^{2/3}\log^{4/3}T\right)$
Future?		

Future work (strongly observable graph)

	Stochastic	Adversarial
Wu et al. (2015)	$O(\alpha \log T / \Delta)$ α is the independence number	
Alon et al. (2015)		$\tilde{O}(\sqrt{\alpha T})$
Erez et al.,(2021) undirected graph with self-loops	$O(\theta \log^4 T / \Delta)$ $\theta \leq \alpha$ is the clique covering number	$\tilde{O}(\sqrt{\theta T})$
Ito et al. (2022)	$O(\alpha \log^3 T / \Delta)$	$\tilde{O}(\sqrt{\alpha T})$
Rouyer et al. (2022) with self-loops	$O(\alpha \log^2 T / \Delta)$	$\tilde{O}(\sqrt{\alpha T})$
Future?		

References

- Alon, Noga, et al. “Online learning with feedback graphs: Beyond bandits.” *Conference on Learning Theory*. PMLR, 2015.
- Wu, Yifan, András György, and Csaba Szepesvári. “Online learning with Gaussian payoffs and side observations.” *Advances in Neural Information Processing Systems* 28 (2015).
- Chen, Houshuang, et al. "Understanding Bandits with Graph Feedback." Advances in Neural Information Processing Systems 34 (2021): 24659-24669.
- Rouyer, Chloé, et al. "A Near-Optimal Best-of-Both-Worlds Algorithm for Online Learning with Feedback Graphs." *arXiv preprint arXiv:2206.00557* (2022).
- Ito, Shinji, Taira Tsuchiya, and Junya Honda. "Nearly Optimal Best-of-Both-Worlds Algorithms for Online Learning with Feedback Graphs." *arXiv preprint arXiv:2206.00873* (2022).
- Erez, Liad, and Tomer Koren. "Best-of-all-worlds bounds for online learning with feedback graphs." *Advances in Neural Information Processing Systems* (2021).