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

Accepted in 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?

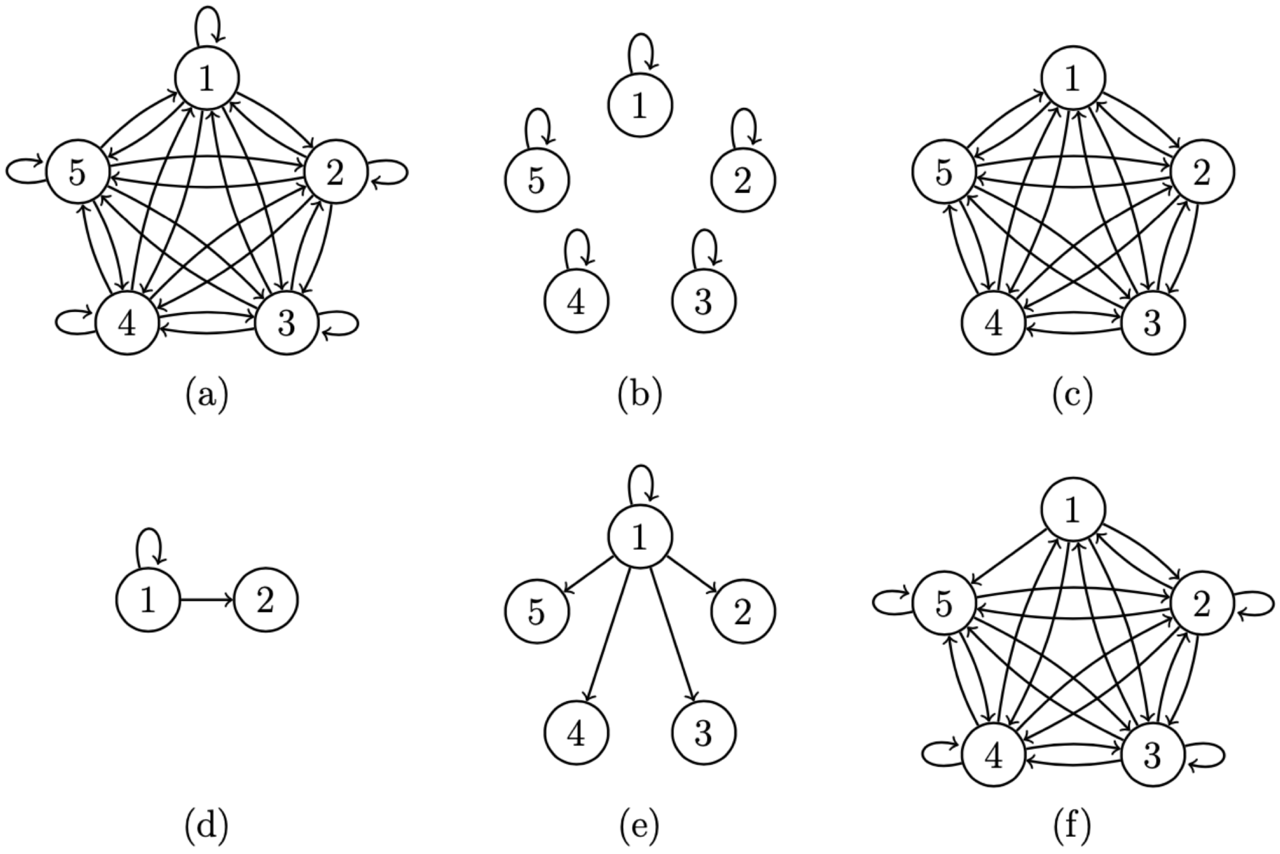
Exploitation V.S. Exploration

General graph feedback

$G = (V, E)$ is the feedback graph

$V = \{1,2, \ldots, K\}$ is the arm set

$E = \{(i, j)\}$ is the directed edge set



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 is rewarded by $r_t(I_t)$

Reward type

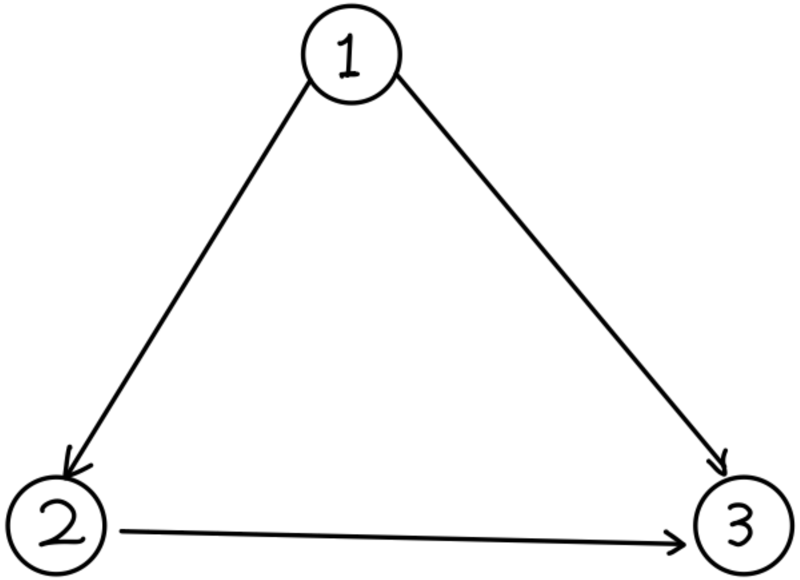
- Stochastic reward
  - $r_t(i)$ is drawn independently from a fixed distribution
  - $\mathbb{E}[r_t(i)] = \mu_i$
  - aim to minimize the regret

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

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

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

Observability



The agent cannot determine which arm is optimal and will suffer $O(T)$ regret.

We consider observable graphs, i.e., $N^{in}(i) \neq \emptyset, \forall i$.

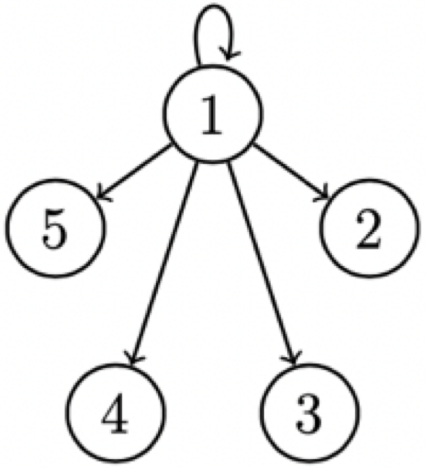
Previous results

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

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 for stochastic setting



Explore: arm 1 (dominating arm set D)
Exploit: arm 1,2,3,4,5 -> 2,3,4 -> 3,4->3

- **Explore-then-commit (ETC) strategy:**
- Select arm 1 until all sub-optimal arms are identified and then focus on the optimal one
- Each arm i need to be observed for $O(\log T/\Delta_i^2)$ times before we identify $\mu_i < \mu_{i^*}$
- $O(|D| \log T/\Delta^2)$

What's wrong if the environment is actually adversarial?

- hope to get $O\left(T^{\frac{2}{3}}\right)$ regret

- ETC would fail in adversarial setting: $O(T)$
  - the optimal arm changes with the horizon

Challenge

- to optimize in the stochastic setting
  - explore dominating arms to collect enough observations
- detect whether the environment is adversarial
  - If the detection condition holds, run the optimal algorithm in adv setting (Alon et al. 2015).
  - Guarantee sublinear regret before the detection condition holds.

Algorithm framework

- 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 adv:
    - If true, run (Alon et al. 2015)

How to guarantee regret before detection holds in the adversarial setting?

- ~~Explore-then-commit~~ $\Rightarrow$ simultaneously explore and exploit
- for t=1,2,3,…
  - $p_{t,D}(i) = \frac{1}{|D|}\mathbb{I}\{i \in D\}, p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$ for each arm $i \in V$
  - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma) p_{t,A}(i)$
  - 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 adv:
    - If true, run Exp3.G (Alon et al. 2015)

While optimizing in the stochastic setting

- for t=1,2,3,…
  - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$;
  - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}$
  - $p_t(i) = \gamma p_{t,D}(i)+(1-\gamma)p_{t,A}(i)$
  - 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 (Alon et al. 2015)

Collect observations to detect adversarial

- for t=1,2,3,...
  - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\};$
  - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j\in D\setminus D_A}\frac{x_j}{t}\right) + \frac{x_i}{t}\mathbb{I}\{i \in D\setminus D_A\}$
  - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma)p_{t,A}(i)$
  - 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 adv:
    - If true, run (Alon et al. 2015)

Detect condition

- 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
  - $\widetilde{H}_t(i) = \frac{1}{t}\sum_{s=1}^{t} \tilde{r}_s(i)$
  - $\left|\widetilde{H}_t(i) - \mu_i\right| \leq \mathrm{radius}_t(i) = O(\sqrt{\frac{1}{t\gamma_t}})$ in stochastic setting
  - $\left|\widetilde{H}_t(i) - \frac{1}{t}\sum_{s=1}^{t} r_s(i)\right| \leq \mathrm{radius}_t(i) = O(\sqrt{\frac{1}{t\gamma_t}})$ in adversarial setting

Detect sub-optimal arms

- for t=1,2,3,…
  - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\};$
  - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j\in D\setminus D_A}\frac{x_j}{t}\right) + \frac{x_i}{t}\mathbb{I}\{i \in D\setminus D_A\}$
  - $p_t(i) = \gamma p_{t,D}(i)+(1-\gamma)p_{t,A}(i)$
  - sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$
  - If $\exists i, j \in A$ such that $\widetilde{H}_t(j) - \widetilde{H}_t(i) > O(\text{ radius}_t(i) + \text{ radius}_t(j))$
    - delete arm i from A
  - detect whether the environment is adv:
    - If true, run (Alon et al. 2015)

Detect adversarial

- for t=1,2,3,…
  - $p_{t,A}(i) = \frac{1}{|A|}\mathbb{I}\{i \in A\}$;
  - $p_{t,D}(i) = \frac{1}{|D_A|}\mathbb{I}\{i \in D_A\}\left(1 - \sum_{j \in D \backslash D_A} \frac{x_j}{t}\right) + \frac{x_i}{t}\mathbb{I}\{i \in D \backslash D_A\}$
  - $p_t(i) = \gamma p_{t,D}(i) + (1-\gamma) p_{t,A}(i)$
  - sample $I_t \sim p_t$ and observe $\left(j, r_t(j)\right)$ for each arm $j \in N^{out}(I_t)$

  - If $\exists i, j \in A$ such that $\widetilde{H}_t(j) - \widetilde{H}_t(i) > \mathrm{radius_t}(i) + \mathrm{radius_t}(j)$
    - delete arm i from A

  - If $\exists i \notin A, j \in A$ such that $\widetilde{H}_t(j) - \widetilde{H}_t(i) < \mathrm{radius_t}(i) + \mathrm{radius_t}(j)$
    - a previous deleted arm becomes better-> adversarial
    - run (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 1-τ rounds: let $i^* \in \mathrm{argmax}_i \sum_{t=1}^{\tau} r_t(i)$
  - $i^* \in A_\tau$
  - $H_\tau(i^*) - H_\tau(i) < \widetilde{H}_\tau(i^*) - \widetilde{H}_\tau(i) + O(\mathrm{radius}_\tau(i^*) + \mathrm{radius}_\tau(i)) < O(\mathrm{radius}_\tau(i))$
  - $O(\tau^{2/3})$

- During $(\tau + 1) - T$ rounds: $O(T^{2/3})$

Regret analysis in stochastic setting

$$Reg(T) = \sum_{i\in V} \Delta_i \, T_i(T)$$
$$\leq \sum_{i\in V} \Delta_i \left(\tau_i^D + \text{resample} + \tau_i\right)$$

- τ_i^D : $\max_{j\in N^{out}(i)} \log T / \Delta_j^2$
- τ_i: $\tilde{O}\left(\frac{\log T}{\Delta_i^2}\right)$
- Resample: lower order

Conclusion

	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

- improve the regret dependence on $T, |D|, K$
  - Recently, Ito et al. (2022) provide $O(|D| \log^2 T / \Delta^2)$ in stochastic setting and $O\left(D^{1/3} T^{2/3} \log^{4/3} T\right)$ in adversarial setting

- better results for graphs with strongly observable graph
  - Ito et al. (2022), Rouyer et al. (2022)