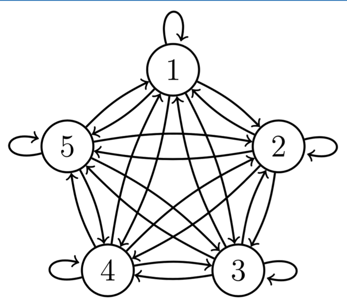
Stochastic Online Learning with Probabilistic Graph Feedback

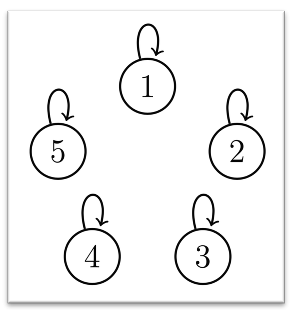
Shuai Li[1]; Wei Chen[2]; Zheng Wen[3]; Kwong-Sak Leung[4]
[1]Shanghai Jiao Tong University; [2]Microsoft Research; [3]DeepMind;
[4]The Chinese University of Hong Kong

Motivation

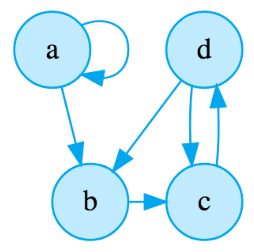
- Full information (complete graph)



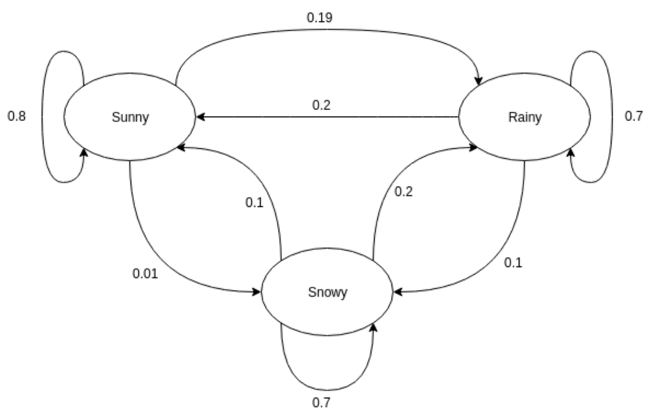
- Bandit feedback (only self-loops)



- Graph feedback can cover full information and bandit settings



- Motivating examples:
  - Recommend island A would infer user's preference of island B
  - Influence spread of seed A would infer influence ability of neighbor B
- Probabilistic graph



Setting

- $V = [K]$ set of actions
- $E \subseteq V \times V$ set of directed edges
- $p: E \to (0,1]$ triggering probabilities
- $\mu = \{\mu_i\}_{i \in V}$ unknown reward distributions
- $\theta = \{\theta_i\}_{i \in V}$ unknown reward means

- At time t,
  - The environment draws a random reward vector
    $$r_t = (r_t(i) : i \in V)$$
    with $r_t(i) \sim \mu_i$ and a random graph $G_t = (V, E_t)$ with $(i,j) \in E$ is active with probability p_{ij}
  - The learner selects action $i_t \in V$ and observes
    One-Step Triggering
    $(j, r_t(j))$ if and only if $(i,j) \in E_t$
    Cascade Triggering
    $(j, r_t(j))$ if and only if there is a path $i \to j$ in G_t
  - The learner receives reward $r_t(i_t)$

- Assumptions
  1. Observability:
     For each action j, there is an edge $(i,j) \in E$ for some i
  2. Reward distributions are of same type:
     $KL(\mu_i, \mu_j)$ is well-defined
  3. Continuity:
     For each μ_i, μ_j and ϵ, there exists μ_i' such that
     $\theta(\mu_i) + \epsilon \le \theta(\mu_i') \le \theta(\mu_i) + 2\epsilon$
     $|KL(\mu_j, \mu_i') - KL(\mu_j, \mu_i)| \le B\epsilon$

- Minimize the expected regret of T rounds
  $$R_\mu(T; G) = T \max_{i \in V} \theta_i - \mathbb{E}\left[\sum_{t=1}^{T} \theta_{i_t}\right]$$

Lower Bounds

- Let $i = 1$ be the best arm
- An algorithm is consistent if $R_\mu(T) = o(T^a)$ for any $a > 0$
- Let p'_{ij} be the probability that there is a directed path from i to j in a random realization of G

One-Step Triggering

$$C(\mu) = \left\{ \begin{array}{l} c \in [0,\infty)^V : \displaystyle\sum_{i \in V^{in}(1)} p_{i1} c_i \ge \frac{1}{KL(\mu_2, \mu_1)} \\ \displaystyle\sum_{i \in V^{in}(j)} p_{ij} c_i \ge \frac{1}{KL(\mu_j, \mu_1)}, \forall j \ne 1 \end{array} \right\}$$

Theorem. For any consistent algorithm, the regret satisfies

$$\underline{\lim}_{T \to \infty} \frac{R_\mu(T)}{\log T} \ge \inf_{c \in C(\mu)} \langle c, \Delta(\mu) \rangle$$

- Recovers existing works for $p \equiv 1$

Cascade Triggering

$$C'(\mu) = \left\{ \begin{array}{l} c \in [0,\infty)^V : \displaystyle\sum_{i \in V^{in}(1)} p'_{i1} c_i \ge \frac{1}{KL(\mu_2, \mu_1)} \\ \displaystyle\sum_{i \in V^{in}(j)} p'_{ij} c_i \ge \frac{1}{KL(\mu_j, \mu_1)}, \forall j \ne 1 \end{array} \right\}$$

Theorem. For any consistent algorithm, the regret satisfies

$$\underline{\lim}_{T \to \infty} \frac{R_\mu(T)}{\log T} \ge \inf_{c \in C'(\mu)} \langle c, \Delta(\mu) \rangle$$

Results

One-Step Triggering

- V^e: the set of exploration nodes that have the largest live probabilities among all incoming edges to some j
- p_i^e: the minimum exploration probability of i

Theorem. The regret satisfies for any $\epsilon > 0$,

$$R(T) = O\left(\log(T) \sum_{i=1}^{K} c_i(\theta, \epsilon) \Delta_i(\theta) + \log(T) \sum_{i \in V^e} \frac{\Delta_i(\theta)}{p_i^e} \right)$$

and

$$\overline{\lim}_{T \to \infty} \frac{R(T)}{\log(T)} \le 4 \cdot \inf_{c \in C(\mu)} \langle c, \Delta(\theta) \rangle$$

holds with probability 1.

- Matched lower bound

Cascade Triggering

- $\hat{V}^e = \left\{ i : p'_{ij} \ge \frac{1}{2} \max_{i'} p'_{i'j} \text{ for some } j \right\}$
  a relaxed version of V^e
- $\hat{p}_i^e = \min\left\{ p'_{ij} : p'_{ij} \ge \frac{1}{2} \max_{i'} p'_{i'j} \text{ for some } j \right\}$
  a relaxed version of p_i^e

Theorem. The regret satisfies for any $\epsilon > 0$,

$$R(T) = O\left(\sum_{i=1}^{K} \Delta_i(\theta) \max_{t \le T} \{ c_i(\theta, \epsilon, \eta(t)) \log t \} + \log(T) \sum_{i \in \hat{V}^e} \frac{\Delta_i(\theta)}{\hat{p}_i^e} \right)$$

Approximation rate of p'_{ij}

and

$$\overline{\lim}_{T \to \infty} \frac{R(T)}{\log(T)} \le 4 \cdot \inf_{c \in C'(\mu)} \langle c, \Delta(\theta) \rangle$$

holds with probability 1.

- Matched lower bound

Algorithms

- $N_i(t)$: selected times of action i
- $M_j(t) = \sum_{i \in V^{in}(j)} N_i(t) p_{ij}$
- $n_{ij}(t)$: observation times of action j when selecting action i
- $m_j(t) = \sum_i n_{ij}(t)$
- $N^e(t)$: number of exploration rounds on θ

One-Step Triggering

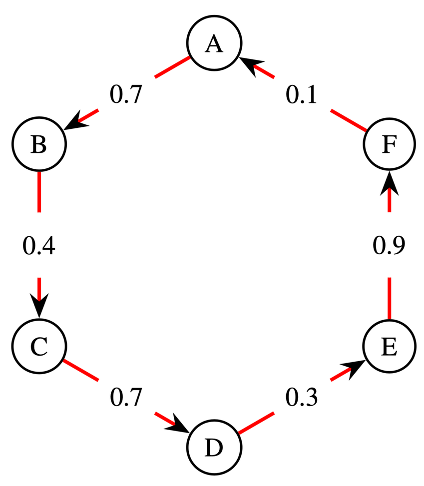
1. $N^e = 0$; $\hat{\theta} = (1, 1, \dots, 1)$
2. For all $t = 1, 2, \dots, T$ do
   Gap between probabilistic graph and realizations
   1) If $m_j < M_j/2$, play $i_t \in \operatorname{argmax}_{i \in V^{in}(j)} p_{ij}$
   Exploitation
   2) Else if $\frac{N(t)}{\log t} \in C(\hat{\theta})$, play $i_t = i_1(\hat{\theta})$
   Forced exploration
   Sublinear auxiliary function
   3) Else if $M_j < 2\beta(N^e)/K$, play $i_t \in V^{in}(j)$ and increase N^e by 1
   Exploration by linear programming
   4) Play $i_t = i$ such that
   $$N_i < 16 c(\hat{\theta}) \log(t)$$
   and increase N^e by 1

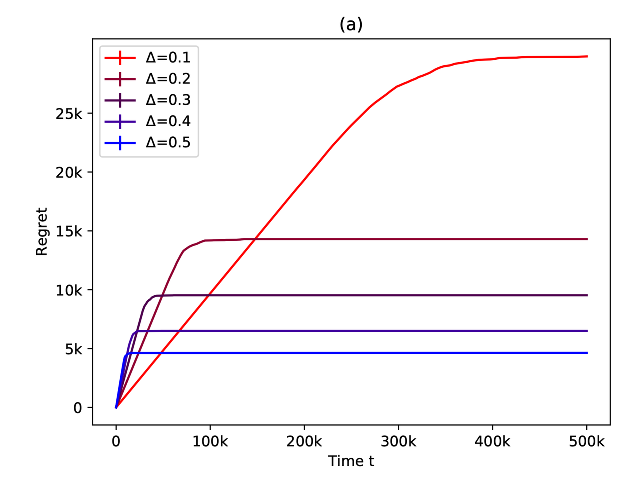
Cascade Triggering

- The computation of p'_{ij} is $\#P$-hard
- Approximate them by $p'_{ij}(t)$
- Run above algorithm on $\tilde{G}_t$ with $p'_{ij}(t)$
- Details omitted

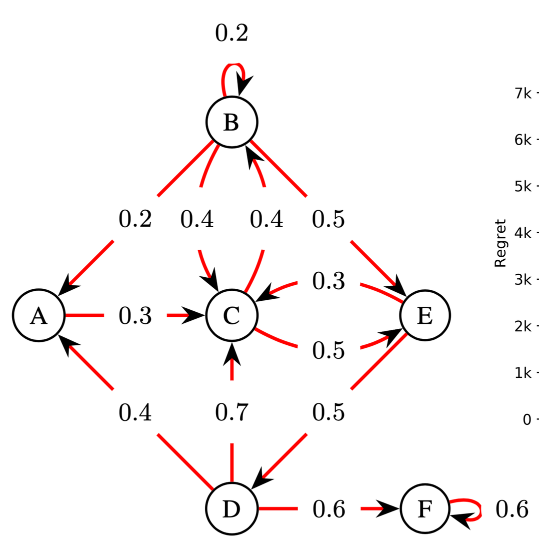
Experiments

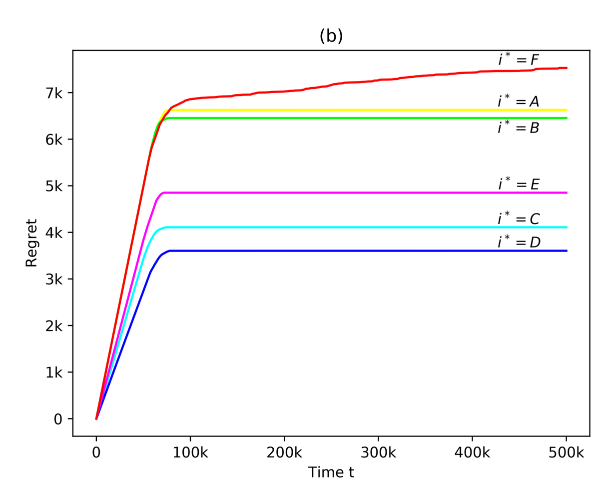
a. Cyclic feedback graph
$$\theta = (0.5 + \Delta, 0.5, \dots, 0.5)$$





b. A random feedback graph
Best arm's reward
$$\theta = (0.6, 0.5, \dots, 0.5)$$





Conclusions

- First work, generalizes existing settings
- Also considers cascade triggering case
- Lower bound
- Upper bound, match with probability 1

Contact

Shuai Li shuaili8@sjtu.edu.cn
Wei Chen weic@microsoft.com
Zheng Wen zhengwen@google.com
Kwong-Sak Leung ksleung@cse.cuhk.edu.hk

Full Paper

Homepage

References

1. Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015a. Online learning with feedback graphs: Beyond bandits. In Conference on Learning Theory, 23–35.
2. Koca´k, T.; Neu, G.; Valko, M.; and Munos, R. 2014. Efficient learning by implicit exploration in bandit problems with side observations. In Advances in Neural Information Processing Systems (NeurIPS), 613–621.
3. Koca´k, T.; Neu, G.; and Valko,M. 2016a. Onlinelearningwitherdő s-ré nyiside- observation graphs. In Uncertainty in Artificial Intelligence (UAI).
4. Liu, F.; Buccapatnam, S.; and Shroff, N. 2018. Information directed sampling for stochastic bandits with graph feedback. In Thirty-Second AAAI Conference on Artificial Intelligence (AAAI).
5. Mannor, S., and Shamir, O. 2011. From bandits to experts: On the value of side- observations. In Advances in Neural Information Processing Systems (NeurIPS), 684–692.
6. Wu, Y.; Gyö rgy, A.; and Szepesvá ri, C. 2015. Online learning with gaussian payoffs and side observations. In Advances in Neural Information Processing Systems (NeurIPS), 1360–1368.