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

Combinatorial Multivariant Multi-Armed Bandits with Applications to Episodic Reinforcement Learning and Beyond

MDP is a Special Case of CMAB

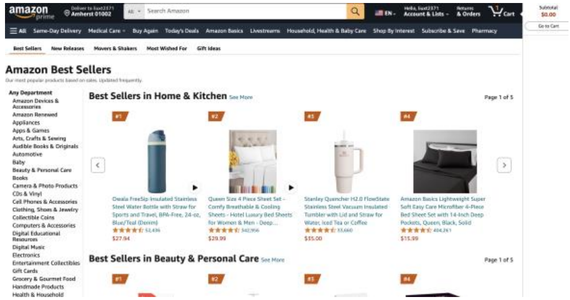
Shuai Li

2024.12.7

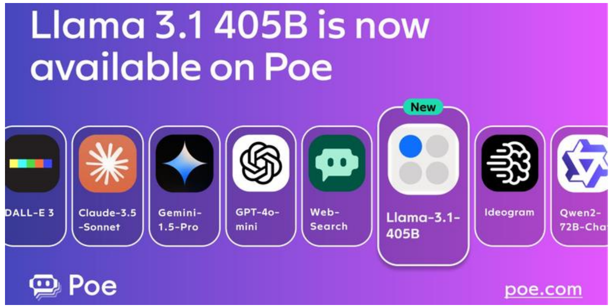
At Fudan University

Making sequential decisions everywhere

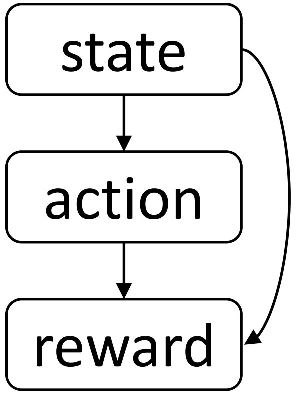
Driving

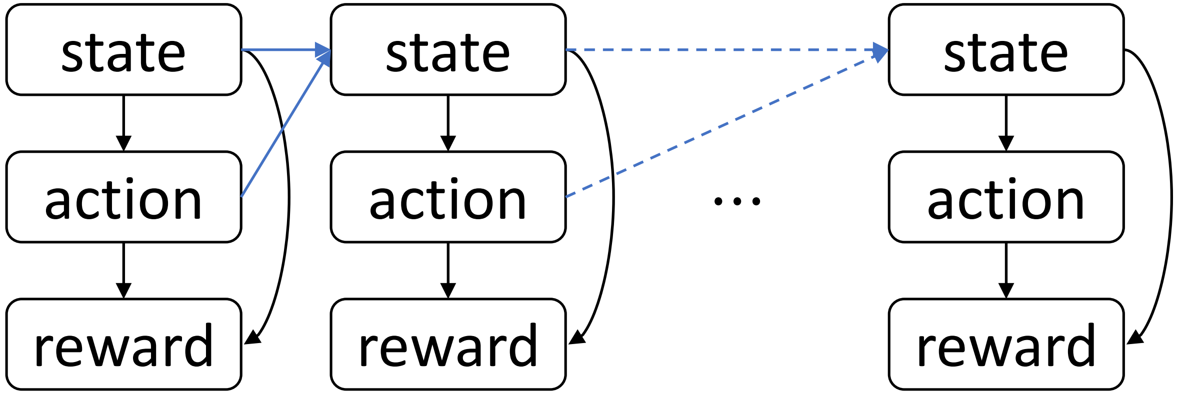


Recommendation



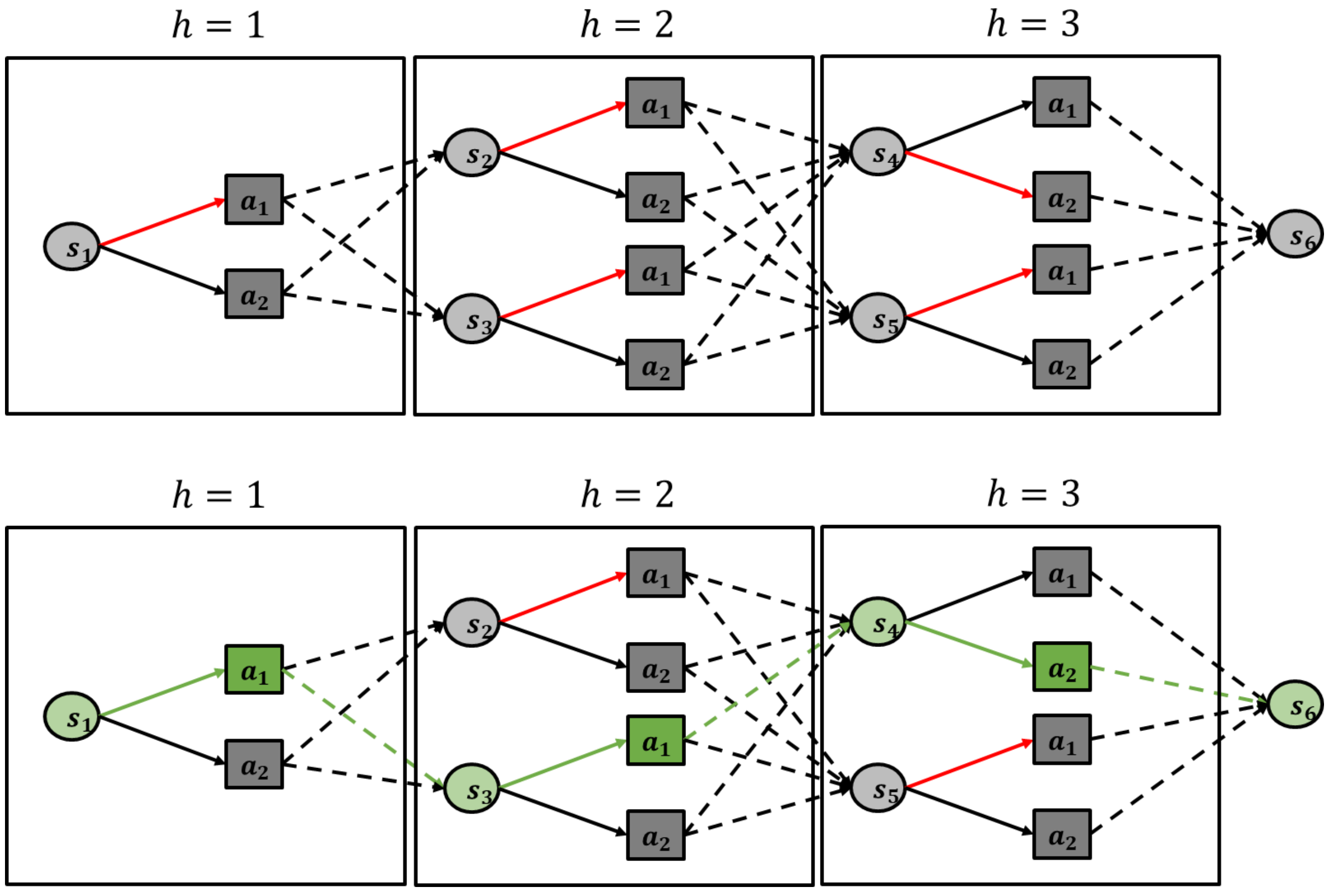
LLM selection





Contextual Bandit $\subseteq$ Markov Decision Process (MDP)

A key observation of MDP



A policy

= An action for each state

= { state-action pairs }

Combinatorial

State-action pairs are triggered to be observed

with triggering

Can MDP be modeled in the framework of Combinatorial MAB?

Multi-armed bandits (MAB)

- A player and m arms (items, products, movies, companies, …)
- Each arm i has a reward distribution P_i with unknown mean μ_i
- In each round $t = 1,2,\ldots$:
  - The agent selects an arm $I_t \in \{1,2,\ldots,m\}$
  - Observes reward $X_t \sim P_{I_t}$
- Objective: Minimize the regret in T rounds

$$\mathrm{Reg}(T) = T \cdot \mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]$$

best arm (i^*)

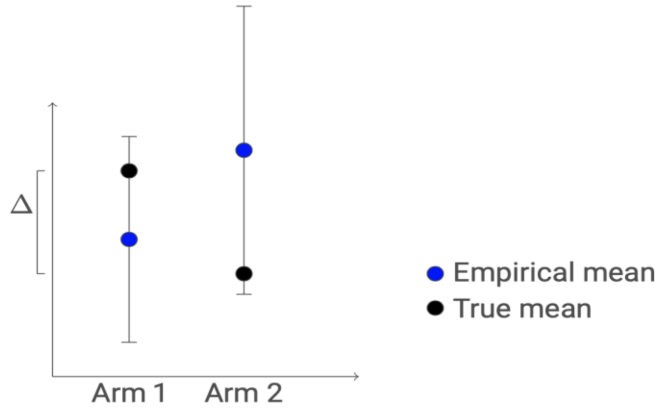
Assume rewards are bounded on [0,1]

Upper confidence bound (UCB) [Auer et al., 2002]

- With high probability $\geq 1-\delta$ By Hoeffding's inequality

$$\mu_i \in \left[\hat{\mu}_i - \sqrt{\frac{\log 1/\delta}{T_i}},\quad \hat{\mu}_i + \sqrt{\frac{\log 1/\delta}{T_i}}\right]$$

Sample mean ($\hat{\mu}_i$) Selection times of arm i (T_i)



- Optimism in face of uncertainty:
  - Believe arms have higher rewards, encourage exploration
- For each round t, select the arm

$$I(t) \in \operatorname{argmax}_{i\in[K]} \left\{\hat{\mu}_i + \sqrt{\frac{\log 1/\delta}{T_i(t)}}\right\}$$

Exploitation ($\hat{\mu}_i$)

Exploration ($\sqrt{\frac{\log 1/\delta}{T_i(t)}}$)

Δ = min gap between best and suboptimal arms

- Regret $\theta(m\log T/\Delta) = \sqrt{mT\log T}$

Combinatorial multi-armed bandits (CMAB)

- A player and m arms
- Each arm i has a reward distribution P_i with unknown mean μ_i
- In each round $t = 1,2, \ldots$:
  - The agent selects an arm set $S_t \subseteq \{1,2, \ldots, m\}$ with size $\leq K$
  - Observes feedback $X_{t,i} \sim P_i$ for each $i \in S_t$
  - Receive reward $R_t(S_t) = \sum_{i \in S_t} X_{t,i}$ with mean $\mu(S_t) = \sum_{i \in S_t} \mu_i$
- Objective: Minimize the regret in T rounds

$$\mathrm{Reg}(T) = T \cdot \mu(S^*) - \mathbb{E}\left[\sum_{t=1}^{T} \mu(S_t)\right]$$

shortest paths, a list of items, influential seed set

semi-bandit feedback

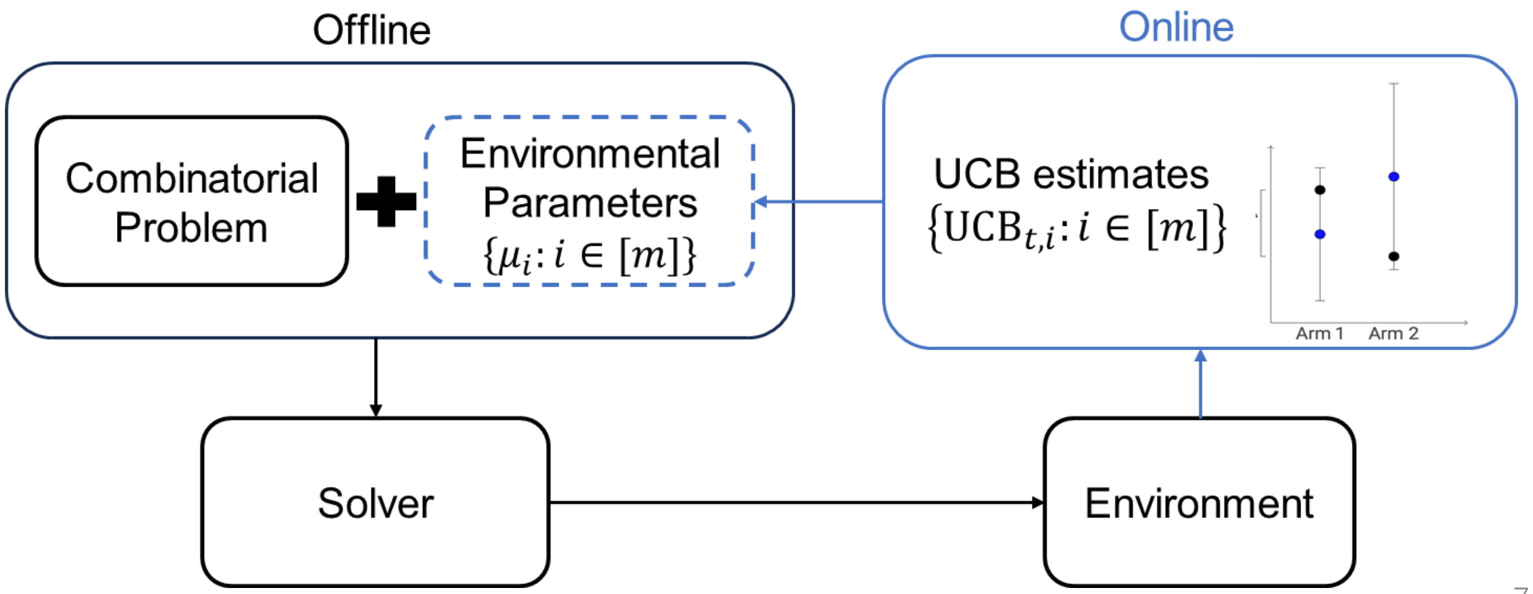
sum reward

best set

Exponential # of actions $\binom{m}{K}$!

Combinatorial UCB [Chen et al., 13]

- In each round $t = 1,2,\ldots:$
  - Compute $\mathrm{UCB}_{t,i} = \hat{\mu}_i + \sqrt{\frac{\log 1/\delta}{T_i(t)}}$ for each arm i
  - Select the action $S = \arg\max_{S:|S|\leq K} \sum_{i\in S} \mathrm{UCB}_{t,i}$
- Regret $\tilde{O}\left(\sqrt{mKT}\right)$



Specialties of MDP: Triggering

- Triggering has been dealt before with CMAB
- $r(S^*;\mu) - r(S_t;\mu)$

$$\le r(S^*;\mathrm{UCB}_t) - r(S_t;\mu)$$
$$\le r(S_t;\mathrm{UCB}_t) - r(S_t;\mu)$$
$$\le \sum_{i \in \tilde{S}_t} p_i(S_t,\mu) \cdot \left|\mathrm{UCB}_{t,i} - \mu_i\right|$$

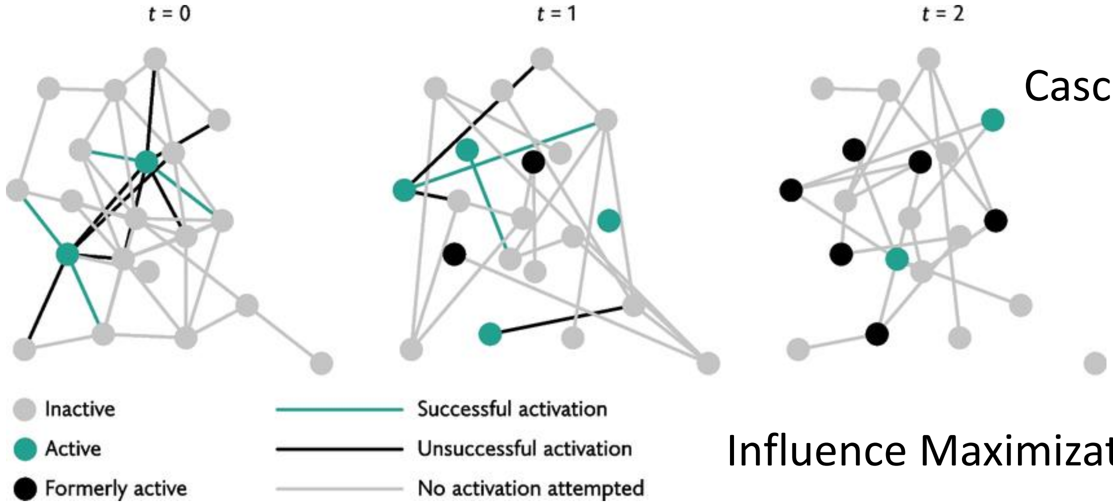
Once observed, the UCB will decrease

Triggering probability

Triggering set

1. Does MDP follows such triggering smoothness ?

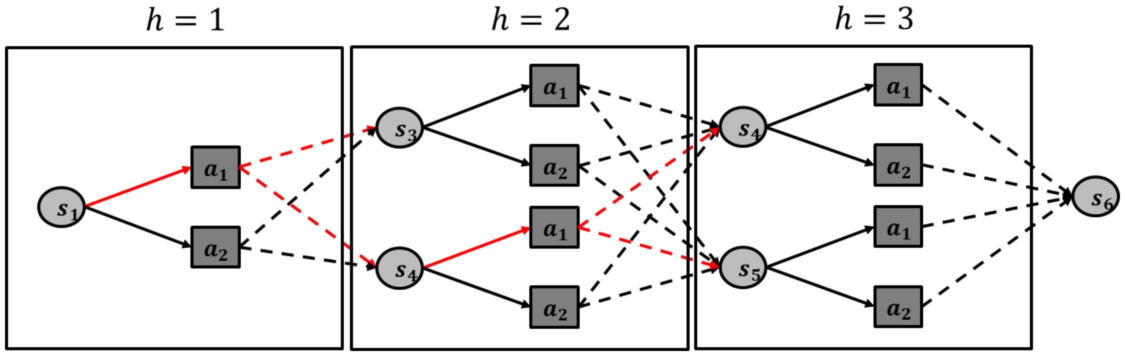
Cascading Click Model



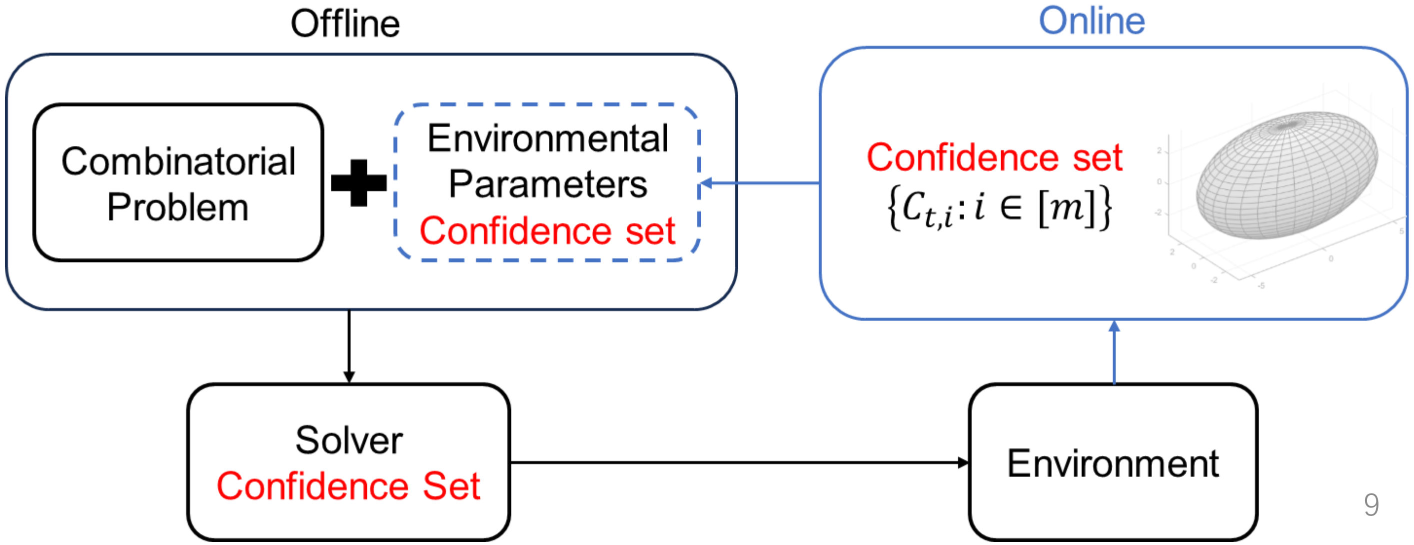
Influence Maximization

Specialties of MDP: Vector-value

- Transition follows categorial distribution



Could also treat transition as S Bernoulli variables
But with worse concentration



2. Does there exists a solver based on confidence set for MDP ?

Triggering smoothness & concentration

- Lemma. Episodic MDP satisfies

$$|V_1(s_1;\tilde{p},\pi) - V_1(s_1;p,\pi)|$$

$$\leq \sum_{s,a,h} q(s,a,h;p,\pi) \left| \left(\tilde{p}(\cdot\,|s,a,h) - p(\cdot\,|s,a,h)\right)^\top V_{h+1}(\cdot\,|\tilde{p},\pi) \right|$$

visitation probability

- (Bound 1) $\leq H \sum_{s,a,h} q(s,a,h;p,\pi) \|\tilde{p}(\cdot\,|s,a,h) - p(\cdot\,|s,a,h)\|_1$

- Concentration

$$\mathcal{C}_t = \left\{ \tilde{p} \in \Delta_S : \|\tilde{p}(\cdot\,|s,a,h) - \hat{p}_t(\cdot\,|s,a,h)\|_1 \leq \sqrt{\frac{2S\log(1/\delta)}{N_t(s,a,h)}} \right\}$$

empirical transition

save $\sqrt{S}$ compared to independent Bernoulli r.v.

# of times visiting (s,a,h)

Offline solver: Extended value iteration

- $(\pi_t, \tilde{p}_t) = \text{argmax}_{\pi, \tilde{p} \in \mathcal{C}_t} V_1(s_1; \tilde{p}, \pi)$
- $h = H, H-1, \ldots, 1$
  - $\tilde{p}_t(\cdot\,|s,a,h) = \text{argmax}_{\tilde{p} \in \mathcal{C}_t} \tilde{p}(\cdot)^\top \bar{V}_{t,h+1}(\cdot)$
  - $Q_t(s,a,h) = r(s,a,h) + \tilde{p}_t(\cdot\,|s,a,h)^\top \bar{V}_{t,h+1}(\cdot)$
  - $\pi_t(s;h) = \text{argmax}_a Q_t(s,a,h)$ and $\bar{V}_{t,h}(s) = \max_a Q_t(s,a,h)$
- Linear problem over a convex polytope, solvable in $O(S^2 A)$
- Regret $\tilde{O}\left(\sqrt{H^4 S^2 A T}\right)$

not optimal !
$\tilde{O}(\sqrt{HS})$- worse than SOTA

Tighter smoothness & concentration

- Lemma. Episodic MDP satisfies $|V_1(s_1;\tilde{p},\pi) - V_1(s_1;p,\pi)|$

$$\le \sum_{s,a,h} q(s,a,h;p,\pi)\left|\left(\tilde{p}(\cdot\mid s,a,h) - p(\cdot\mid s,a,h)\right)^\top V_{h+1}(\cdot\mid\tilde{p},\pi)\right|$$

- $\mathcal{C}_t = \left\{\tilde{p} \in \Delta_S : \left(\tilde{p}(\cdot\mid s,a,h) - \hat{p}_t(\cdot\mid s,a,h)\right)^\top V_{h+1}(\cdot\mid\tilde{p},\pi) \le \tilde{O}\left(\sqrt{\frac{\mathrm{Var}_{p(\cdot|s,a,h)}[V^*_{h+1}(\cdot)]}{N_t(s,a,h)}}\right)\right\}$

unknown

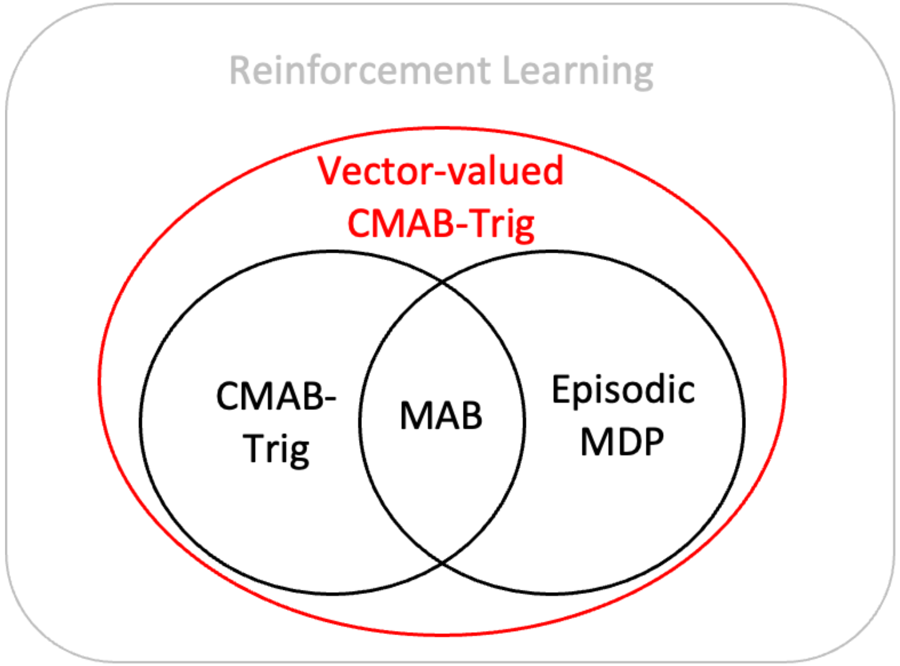
- $\le \phi_t(s,a,h) = \tilde{O}\left(\sqrt{\frac{\mathrm{Var}_{\hat{p}_{t-1}(\cdot|s,a,h)}[\overline{V}_{t,h+1}(\cdot)]}{N_t(s,a,h)}} + \sqrt{\frac{\mathbb{E}_{\hat{p}_{t-1}(\cdot|s,a,h)}[\overline{V}_{t,h+1}(\cdot) - \underline{V}_{t,h+1}(\cdot)]^2}{N_t(s,a,h)}} + \frac{5H}{N_t(s,a,h)}\right)$

Offline solver: Optimistic value iteration

- $(\pi_t, \tilde{p}_t) = \text{argmax}_{\pi, \tilde{p} \in \mathcal{C}_t} V_1(s_1; \tilde{p}, \pi)$
- $h = H, H-1, \ldots, 1$
  - $\tilde{p}_t(\cdot\,|s,a,h) = \text{argmax}_{\tilde{p} \in \mathcal{C}_t} \tilde{p}(\cdot)^\top \bar{V}_{t,h+1}(\cdot)$
  - $Q_t(s,a,h) = \boxed{r(s,a,h) + \phi_t(s,a,h)} + \tilde{p}_t(\cdot\,|s,a,h)^\top \bar{V}_{t,h+1}(\cdot)$
  - $\pi_t(s;h) = \text{argmax}_a Q_t(s,a,h)$ and $\bar{V}_{t,h}(s) = \max_a Q_t(s,a,h)$

Result

- Regret $O\left(\sqrt{H^3SAT\log(SAHT)} + H^3S^2A\log^{3/2}(SAHT)\right)$
  - Match lower bound $\Omega\left(\sqrt{H^3SAT}\right)$ up to log factors
  - Save $O\left(\log^{5/2}(SAHT)\right)$ factor for $O(\sqrt{T})$ term compared to [Zanette and Brunskill, 19]
  - As a by-product, this work could derive gap-dependent bound naturally
    [Simchowitz and Jamieson, 19] use a very complicated analysis to derive gap-free bound from gap-dependent bound



Beyond MDP

Generalization of smoothness & confidence

- Lemma. Episodic MDP satisfies

$$|V_1(s_1;\tilde{p},\pi) - V_1(s_1;p,\pi)| \le \sum_{s,a,h} q(s,a,h;p,\pi)\left|\left(\tilde{p}(\cdot\,|s,a,h) - p(\cdot\,|s,a,h)\right)^\top V_{h+1}(\cdot\,|\tilde{p},\pi)\right|$$

- Assumption (Smoothness Condition).

$$|r(\pi;\tilde{\mu}) - r(\pi;\mu)| \le \sum_{i\in[m]} q(i;\mu,\pi)\cdot\left|\left(\tilde{\mu}_i(\cdot) - \mu_i(\cdot)\right)^\top w_i(\cdot\,|\tilde{\mu},\pi)\right|$$

triggering probability

weight $\in [0, w]$
depend on policy

- Confidence region

$$\mathcal{C}(\pi) = \left\{\tilde{\mu} \in [0,1]^{m\times d}: \left|\left(\tilde{\mu}_i(\cdot) - \hat{\mu}_i(\cdot)\right)^\top w_i(\cdot\,|\tilde{\mu},\pi)\right| \le F_i\sqrt{\frac{1}{N(i)}} + \frac{\bar{I}}{N(i)}, \forall i \in [m]\right\}$$

where $\sum_{i\in[m]} q(i;\mu,\pi)F_i^2 \le \bar{F}$

Generalization of solver

- Assumption (Offline Oracle).
  Input: Confidence region $\mathcal{C}$ defined on policy
  Output: Action-parameter pair $(\tilde{\pi}, \tilde{\mu}) = \tilde{\mathcal{O}}(\mathcal{C})$ s.t.
  - $\tilde{\pi} \in \Pi, \tilde{\mu} \in \mathcal{C}(\tilde{\pi})$
  - is an α-approximation, i.e.,
    $$r(\tilde{\pi}; \tilde{\mu}) \geq \alpha \cdot \max_{\pi, \mu \in \mathcal{C}(\pi)} r(\pi; \mu)$$

- Objective: Minimize α-Regret $\mathbb{E}[\sum_t \alpha \cdot r(\pi^*; \mu) - r(\pi_t; \mu)]$

Result

- **Theorem**. CUCB-MT achieves an α-approximate regret of
$$O\left(\sqrt{m(\bar{F}+\bar{G})T}+m(\bar{I}+\bar{J})\log(KT)\right)$$
  - **(Concentration 1)** $\mu \in \mathcal{C}_t(\pi^*)$
  - **(Concentration 2)**

$$\left|\left(\mu_i(\cdot)-\hat{\mu}_i(\cdot)\right)^\top\left(w_i(\cdot\,|\tilde{\mu}_t,\pi_t)-w_i(\cdot\,|\mu,\pi^*)\right)\right| \le G_i\sqrt{\frac{1}{N_t(i)}}+\frac{\bar{J}}{N_t(i)}$$

for $(\pi_t,\tilde{\mu}_t)=\tilde{\mathcal{O}}(\mathcal{C}_t)$
where $\sum_{i\in[m]} q(i;\mu,\pi)G_i^2 \le \bar{G}$

Analysis

- Regret decomposition + CMAB-T analysis (e.g., triggering probability equivalence, reverse amortization, regret allocation)

$$\Delta_{\pi_t} = \alpha \cdot r(\pi^*; \boldsymbol{\mu}) - r(\pi_t; \boldsymbol{\mu})$$

$$\overset{\text{(joint oracle)}}{\leq} r(\pi_t; \tilde{\boldsymbol{\mu}}_t) - r(\pi_t; \boldsymbol{\mu})$$

$$\overset{\text{(1-norm MTPM)}}{\leq} \sum_{i \in [m]} q_i^{\boldsymbol{\mu}, \pi_t} \left| (\tilde{\boldsymbol{\mu}}_{t,i} - \boldsymbol{\mu}_i)^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}, \pi_t} \right|$$

$$\leq \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \left| (\tilde{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\mu}}_i)^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}, \pi_t} \right| + q_i^{\boldsymbol{\mu}, \pi_t} \left| (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i})^\top \boldsymbol{w}_i^{\boldsymbol{\mu}, \pi^*} \right|$$

$$+ q_i^{\boldsymbol{\mu}, \pi_t} \left| (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i})^\top (\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}, \pi_t} - \boldsymbol{w}_i^{\boldsymbol{\mu}, \pi^*}) \right| + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu}, \pi_t} \bar{w} d$$

$$\overset{\text{(concentration)}}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \sqrt{\frac{(2F_{t,i} + G_{t,i})^2}{N_{t-1,i}}} + q_i^{\boldsymbol{\mu}, \pi_t} \frac{2I_{t,i} + J_{t,i}}{N_{t-1,i}} + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu}, \pi_t} \bar{w} d$$

Application: Probabilistic Maximum Coverage

- Probabilistic maximum coverage (PMC)
  - Weighted bipartite graph $G = (U, V, E, p)$
  - Each vertex $u \in U$ independently try to cover its neighbor $v \in V$
- Probabilistic maximum coverage for goods distribution (PMC-GD)
  - Weighted bipartite graph $G = (U, V, E, p)$
  - Each vertex $u \in U$ will cover one of its neighbor $v \in V$ and $\sum_v p_{u,v} \leq 1$
- CUCB-MT achieves $(1 - 1/e)$-regret $\tilde{O}(\sqrt{K|U||V|T})$
  - K is the seed set size
  - Improve over existing work [Wang & Chen, 17] by a factor of $\sqrt{|V|}$

Thanks!
&
Questions?

Shuai Li

- Associate Professor at John Hopcroft Center Shanghai Jiao Tong University
- Research interests: Bandit/RL algorithms
- Personal website: http://shuaili8.github.io/