Decentralized Cooperative Reinforcement Learning with Hierarchical Information Structure

Hsu Kao
University of Michigan
hsukao@umich.edu
Joint work with Chen-Yu Wei and Vijay Subramanian

ALT, March 29, 2022

Tackling Non-stationarity

- One of major challenges in MARL: non-stationarity
- Solutions:
  - Communication

    → communication overhead, privacy loss
  - Coordination (common information approach)

    → computation overhead, privacy loss, may require shared randomness

Hierarchical Decision Making

- Sequential decision making
  - Decisions from agents that act before are known (hierarchical information structure for *actions*)
  - Two agents: Stackelberg game-like setting with leader/follower action spaces $[A]/[B]$ and common objective
- Widely applicable
  - Cognitive radio: primary user/secondary user
  - Organizations (corporate/government): high/low levels

Can we design hierarchical cooperative MARL where the leader is completely uninformed of the follower's actions/policies? (i.e., joint exploration with no communication/coordination)

Hierarchical Cooperative Bandit

- Input: A, B, $\{\mu_{a,b}\}_{a\in[A],b\in[B]}$ (unknown)

for $t = 1, \dots, T$ **do**
  Leader chooses $a_t \in [A]$
  After receiving a_t, follower chooses $b_t \in [B]$
  Both agents receive reward $r_t = \mu_{a_t,b_t} + \underbrace{\text{noise}}$
end

Only has to be zero mean
(potentially agent-dep)

- Goal: minimize $\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(\max_{a,b} \mu_{a,b} - \mu_{a_t,b_t} \right)$
  - b_t not observable to leader, introducing asymmetry
  - a_t observable to follower, creating "hierarchy"

Hierarchical Cooperative MDP

- Input: episodic MDP S, A, B, $\boxed{P,\ R}$ (unknown)

```
for t = 1, … , T do
    for h = 1, … , H do
        Leader chooses a_{t,h} ∈ [A]
        After receiving a_{t,h}, follower chooses b_{t,h} ∈ [B]
        Both agents receive reward r_{t,h} = R(s_{t,h}, a_{t,h}, b_{t,h}) + noise
                    observe next state s_{t+1,h} ~ P(· | s_{t,h}, a_{t,h}, b_{t,h})
    end
end
```

  - $\pi_h^1: [S] \to [A]$, $\pi_h^2: [S] \times [A] \to [B]$; $\pi = \left(\{\pi_h^1\}_{h=1}^H, \{\pi_h^2\}_{h=1}^H\right)$
  - $V^\pi(s) := \mathbb{E}[R(s_h, a_h, b_h)|s_1 = s, \pi]$

- Goal: minimize $\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^T \left(\max_\pi V^\pi(s_{t,1}) - \sum_{h=1}^H r_{t,h}\right)\right]$

Hierarchical Cooperative MARL and Literature

- Key idea
  - Follower uses no-regret algorithm given a leader's action
  - Leader adopts inflated bonus aligned with follower's regret bound, allowing follower to converge to optimum first
- Settings: bandit, MDP
- Literature:
  - CI approach: single-agent with action space $\mathcal{A} \times \mathcal{B}$, gives lower bound (e.g. [Chang21])
  - Hierarchical structure: idea of upper bonus = lower regret
    - Modified UCT [Coqueling07] (MC tree search)
    - Stochastic corral [Arora21] (model selection)
  - MAMAB: most need communication/coordination
  - Markov game: different equilibria/convergence concepts

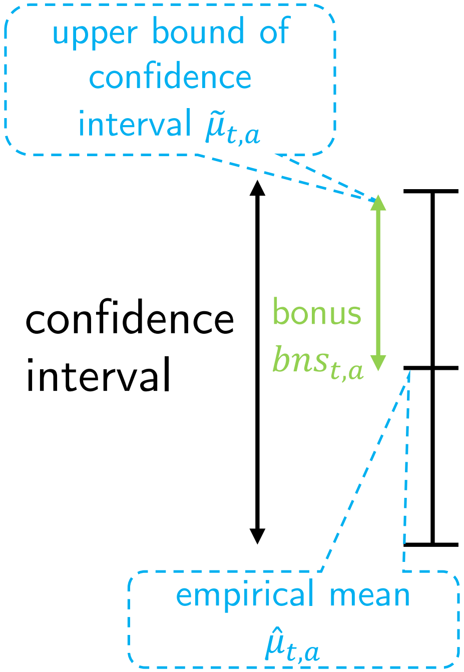
Review: Single MAB and UCB Algorithm

Agent chooses $a_t \in [A]$, then receives $r_t = \mu_{a_t} + \mathrm{noise}$

- UCB principle [Agarwal95, Auer02]:

optimistic value = empirical value + bonus $\geq$ true (/optimal) value w.h.p.

$$\tilde{\mu}_{t,a} = \underbrace{\frac{\sum_{\tau=1}^{t-1} \mathbb{I}\{a_t = a\} r_\tau}{n_t(a)}}_{\hat{\mu}_{t,a}\text{: empirical mean of } a \text{ until } t-1} + c \underbrace{\sqrt{\frac{\log t}{n_t(a)}}}_{bns_{t,a}}$$



$bns_{t,a}$ is s.t. (1) $\tilde{\mu}_{t,a} \geq \mu_a$ w.h.p.
$\Leftrightarrow bns_{t,a} \geq \hat{\mu}_{t,a} - \mu_a$
(2) $\sum_t bns_{t,a} = o(T)$

$$a_t \in \operatorname{argmax}_a \tilde{\mu}_{t,a} \Rightarrow \mathrm{Reg}(T) \approx \text{concentration bd} + \sum_t bns_{t,a_t} \lesssim \sqrt{AT}$$

UCB Principle for Leader

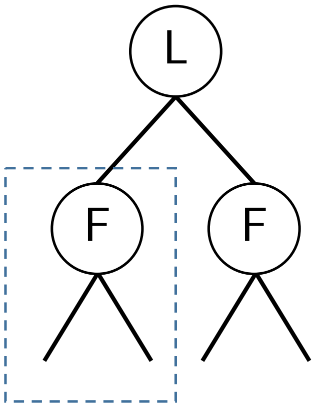
$$bns_{t,a} \geq \text{true value of } a - \text{ empirical value of } a$$

$$= \max_b \mu_{a,b} - \frac{\sum_{\tau=1}^{t-1} \mathbb{I}\{a_t = a\} r_\tau}{n_t(a)}$$

$$= \frac{1}{n_t(a)} \sum_{\tau=1}^{t-1} \mathbb{I}\{a_t = a\} \left(\max_b \mu_{a,b} - r_\tau \right)$$

$$= \text{follower's average regret under } a$$

$$= \frac{\mathrm{Reg}_a(n_t(a))}{n_t(a)}$$



standard
MAB

Hierarchical Bandit Algorithm

- Follower runs no-regret algo: $\mathrm{Reg}_a(\tau) \lesssim \sqrt{B\tau \log \tau}$
- Leader runs UCB with bonus inflated by $\sqrt{B}$

$$bns_{t,a} \geq \frac{\mathrm{Reg}_a\big(n_t(a)\big)}{n_t(a)} \approx \sqrt{\frac{B \log t}{n_t(a)}}$$

$\Rightarrow \mathrm{Reg}(T) \approx \text{concentration bd} + \sum_t bns_{t,a_t} \lesssim \sqrt{ABT}$

Review: Single MDP and UCBVI/UCB-Q

Agent chooses $a_{t,h} \in [A]$, receives reward $r_{t,h} = R(s_{t,h}, a_{t,h}) + \text{noise}$
observe next state $s_{t+1,h} \sim P(\cdot \mid s_{t,h}, a_{t,h})$

- No-regret algos: UCBVI [Azar17], UCB-Q [Jin18]

optimistic value = empirical value + bonus $\geq$ optimal value   w.h.p.

- $\tilde{Q}_{t,h}(s,a)$: adding bonus in Bellman updates
- Action ⇨ Policy (π_t optimal w.r.t. $\tilde{Q}_t$)

Bonus is s.t. (1) $\tilde{Q}_{t,h}(s,a) \geq Q_h^*(s,a)$ w.h.p.
(2) $\sum_t \left[\tilde{Q}_{t,h}(s_{t,h}, a_{t,h}) - Q_h^{\pi_t}(s_{t,h}, a_{t,h})\right] = o(T)$

Two-Player Hierarchical MDP

- Leader: $\tilde{Q}^1_{t,h}(s,a) \geq Q^*_h(s,a) \triangleq \max_b Q^*_h(s,a,b)$
  - π^1_t optimal w.r.t. $\tilde{Q}^1_t$
- Follower: $\tilde{Q}^2_{t,h}(s,a,b) \geq Q^*_h(s,a,b)$
  - π^2_t optimal w.r.t. $\tilde{Q}^2_t$
- Problem:
  - Follower is more informed
  - Optimism: explore $\left(a_{t,h}, b_{t,h}\right) = \operatorname{argmax}_{a,b} Q^2_{t,h}\left(s_{t,h}, a, b\right)$
  - Leader doesn't know

How to perform **joint exploration** without communication?

The Key Property

$$\mathrm{Reg}(T) \approx \underbrace{\textstyle\sum_{t,h}\left[V_h^*(s_{t,h}) - Q_h^*(s_{t,h}, a_{t,h})\right]}_{\mathrm{Reg}^1(T)} + \underbrace{\textstyle\sum_{t,h}\left[Q_h^*(s_{t,h}, a_{t,h}) - Q_h^*(s_{t,h}, a_{t,h}, b_{t,h})\right]}_{\mathrm{Reg}^2(T)}$$

Leader's optimism?

?

$\le$

$\le$

$\tilde{Q}^1_{t,h}(s_{t,h}, a_{t,h})$

$\tilde{Q}^2_{t,h}(s_{t,h}, a_{t,h}, b_{t,h})$

Reg^2 depends on π^1

Follower's optimism w.r.t. π^*

Follower's optimism given π^1

The Key Property

$$\mathrm{Reg}(T) \approx \underbrace{\textstyle\sum_{t,h}\left[V_h^*(s_{t,h}) - Q_h^*(s_{t,h}, a_{t,h})\right]}_{\le\ \mathrm{Reg}^1(T)} + \underbrace{\textstyle\sum_{t,h}\left[Q_h^*(s_{t,h}, a_{t,h}) - Q_h^*(s_{t,h}, a_{t,h}, b_{t,h})\right]}_{\le\ \mathrm{Reg}^2(T)}$$

$$\boxed{\tilde{Q}^1_{t,h}(s_{t,h}, a_{t,h})} \ \ge \ \boxed{\tilde{Q}^2_{t,h}(s_{t,h}, a_{t,h}, b_{t,h})}$$

- Make leader more optimistic than follower!
  - Leader runs UCB-Q with bonus inflated by $\sqrt{SB}$
  - Follower runs UCBVI

UCB-Q + UCBVI for Hierarchical MDP

Leader: UCB-Q

U1 updates Q/V functions ($\approx$ UCB-H update rule):
$V^1_{H+1}(\cdot) \leftarrow 0.$
for $h = 1, \ldots, H$ **do**
$Q^1_h(s_h, a_h) \leftarrow (1-\alpha_\tau) Q^1_h(s_h, a_h) + \alpha_\tau \left(r_h + V^1_{h+1}(s_{h+1}) + \mathsf{bns}^1_\tau \right)$
$V^1_h(s_h) \leftarrow \min\{\max_a Q^1_h(s_h, a),\ H\}$
where $\tau = n_h(s_h, a_h)$.

U2 updates Q/V functions ($\approx$ UCBVI update rule):
Let $\hat{P}_h(s'|s,a,b) = \frac{n_h(s,a,b,s')}{n_h(s,a,b)}$ and $\hat{R}_h(s,a,b) = \frac{\theta_h(s,a,b)}{n_h(s,a,b)}$ $\forall h, s, a, b, s'$.
(if $n_h(s,a,b) = 0$, set $\hat{P}_h(s'|s,a,b) = \frac{1}{|S|}$ and $\hat{R}_h(s,a,b) = 0$).
$V^2_{H+1}(\cdot) \leftarrow 0.$
for $h = H, \ldots, 1$ **do**
for *all* s, a, b **do**
$Q^2_h(s,a,b) \leftarrow \min\left\{ \hat{R}_h(s,a,b) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a,b)} \left[V^2_{h+1}(s') \right] + \mathsf{bns}^2_\tau,\ \ Q^2_h(s,a,b) \right\}$
$V^2_h(s) \leftarrow \max_{a,b} Q^2_h(s,a,b)$
where $\tau = n_h(s,a,b)$.

Follower: UCBVI

– Why UCB-Q+UCBVI: UCB-Q/UCBVI shrinks confidence set slower/faster

$\Rightarrow \mathrm{Reg}(T) \lesssim \sqrt{H^7 S^2 ABT}$ lower bound $= \Omega\left(\sqrt{H^3 SABT}\right)$

Conclusion

- Achieved no communication/coordination joint exploration for cooperative bandits/MDPs with hierarchical information structure
  - Leader's exploration bonus $\Leftrightarrow$ follower's regret bound
  - Make leader more optimistic than follower
- Future directions:
  - Closing current/lower bounds ($H^2\sqrt{S}$ factor)
  - General-sum Markov games with hierarchical structure
  - Low-regret learning with other information asymmetry (e.g. reward) or information structure