Oct. 5, 23

Lec 8: LSVI-UCB in linear MDP

- RL problem $\quad M = (S, A, \{r_h\}_{h\in[H]}, \{P_h\}_{h\in[H]})$

interaction protocol

For episode $k = 1, \dots, K$:

- the adversary picks an initial state x_1^k
- the learner plays π^k (based on all information the learner has observed so far)

After K episodes, the regret of the learner:

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^*(x_1^k) - V_1^{\pi^k}(x_1^k) \right]$$

recall:

$$V_1^\pi(x_1) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} r_h(x_h, a_h) \right] \text{ where } (x_1, a_1, \dots, x_H, a_H) \sim (\{P_h\}_{h\in[H]}, \pi)$$

$$V_1^*(x_1) = \max_\pi V_1^\pi(x_1)$$

Linear MDP

$$\exists \ \phi: \ S \times A \to \mathbb{R}^d$$

$$\theta_h \in \mathbb{R}^d$$

$$\mu_h(\cdot) = (\mu_h^{(1)}(\cdot), \dots, \mu_h^{(d)}(\cdot))$$

s.t.

$$\begin{cases} P_h(x' \mid x, a) = \phi(x,a)^T \mu_h(x') \\ r_h(x, a) = \phi(x,a)^T \theta_h \end{cases}$$

Chi Jin, Zhuoran Yang, Zhaoran Wang, Michael Jordan.

"Probably efficient RL with linear function approximation." COLT'2020

Least-square Value Iteration w/ UCB (LSVI-UCB)

For episode $k = 1, \dots, K$

- receive initial state x_1^k
- For step $h = H, \dots, 1$ do

$$\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\, \phi(x_h^\tau, a_h^\tau)^T + \lambda I$$

$$w_h^k \leftarrow \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \left[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a) \right]}$$

$$= \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\tau=1}^{k-1} \left(\phi(x_h^\tau, a_h^\tau)^T w - r_h(x_h^\tau, a_h^\tau) - \max_a Q_{h+1}(x_{h+1}^\tau, a) \right)^2$$

$$Q_h^k(\cdot,\cdot) \leftarrow \min\left\{ \phi(\cdot,\cdot)^T w_h^k + \beta \underbrace{\| \phi(\cdot,\cdot) \|_{(\Lambda_h^k)^{-1}}}_{\text{estimation error}},\ H \right\}$$

- encourage to take actions w/ large estimation error $\longrightarrow$ exploration
- if all actions are well explored, the est. error is small $\rightarrow$ exploitation

- for step $h = 1, \dots, H$:

take action $a_h^k \leftarrow \underset{a}{\operatorname{argmax}}\ Q_h(x_h^k, a)$ and observe $x_{h+1}^k \sim P_h(\cdot \mid x_h^k, a_h^k)$

Theorem: If choose $\beta = \tilde{\Theta}(dH)$, then $\text{regret}_{\text{LSVI-UCB}}(K) = \tilde{O}(H^2 d^{3/2} \sqrt{K})$

LSVI-UCB (Jin et al. COLT'20)	$\tilde{O}(\sqrt{d^3 H^4 K})$
Lower bound (Zhou et al. COLT'21)	$\Omega(d\sqrt{H^3 K})$
LSVI-UCB++ (He et al. ICML'23)	$\tilde{O}(d\sqrt{H^3 K})$

$O(\sqrt{dH})$ gap

variance-aware LSVI to shave off $\sqrt{H}$ factor

reference-advantage decomposition $(V^k_{h+1} = V^*_{h+1} + V^k_{h+1} - V^*_{h+1})$ to shave off $\sqrt{d}$ factor

Zhou et al. COLT'20: Nearly minimax optimal RL for linear mixture MDP

He et al., ICML'23: Nearly minimax optimal RL for linear MDP

Note: Q-learning paper of Chi Jin only give lower bound of order $\Omega(\sqrt{dH^3K})$ if we set $d = SA$

Sketch proof of the theorem

- $E_1 = \{ Q_h^k(x,a) \geq Q_h^*(x,a), \ \forall (x,a,h,k) \in S \times A \times [H] \times [K] \}$ (optimism)

- $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_k}(x_h^k)$ (computation of V_h^k, π_k do not use (x_h^k, a_h^k))

  $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k \mid x_h^k, a_h^k] - \delta_{h+1}^k$

$E_2 = \{ \delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta \|\phi_h^k\|_{(\Lambda_h^k)^{-1}} : \forall (k,h) \}$

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^{*}(x_1^k) - V_1^{\pi_k}(x_1^k) \right]$$

$$\overset{E_1}{\leq} \sum_{k=1}^{K} \left[V_1^{k}(x_1^k) - V_1^{\pi_k}(x_1^k) \right]$$

$$= \sum_{k=1}^{K} \delta_1^k \overset{E_2}{\leq} \sum_{k,h} \zeta_h^k + 2\beta \sum_{k,h} \| \phi_h^k \|_{(\Lambda_h^k)^{-1}}$$

- $\{\zeta_h^k\}_{k,h}$ is a martingale difference sequence. and $|\zeta_h^k| \leq 2H$

Azuma-Hoeffding's ineq.

$$\sum_{k,h} \zeta_h^k \leq \sqrt{KH^2 \log(1/\delta)} \quad \text{w.p } 1-\delta$$

- $$\sum_{k=1}^{K} \| \phi_h^k \|^2_{(\Lambda_h^k)^{-1}} \leq 2 \log \frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)} = O(d \log K)$$

recall: $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^T + \lambda I$

- $$\sum_{k,h} \| \phi_h^k \|_{(\Lambda_h^k)^{-1}} \leq \sum_h \sqrt{K \sum \| \phi_h^k \|^2_{(\Lambda_h^k)^{-1}}} \leq H\sqrt{Kd \log K}$$

Sketch proof for $\Pr(E_2)$ is large

- Concentration lemma

$$\forall (k,h): \quad \left\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\left[V_{h+1}^k(x_{h+1}^\tau) - P_h V_{h+1}^k (x_h^\tau, a_h^\tau)\right] \right\|_{(\Lambda_h^k)^{-1}} \leq \tilde{O}(dH)$$

- Fix π,

$$\forall (x,a,h,k): \quad \phi(x,a)^T w_h^k - Q_h^\pi(x,a) \leq P_h (V_{h+1}^k - V_{h+1}^\pi)(x,a) + \underbrace{dH}_{\beta} \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}$$

E_2

E_1

Sketch proof for $\Pr(E_{\#})$ β large

- $$\phi(x,a)^T w_H^k - Q_H^*(x,a) \le \beta \|\phi(x,a)\|_{(\Lambda_H^k)^{-1}}$$

- $$\phi(x,a)^T w_h^k - Q_h^*(x,a) - \underbrace{P_h (V_{h+1}^k - V_{h+1}^*)(x,a)}_{\le 0 \text{ by induction}} \le \beta \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}$$