07/20/23

Linear MDP w/ a simulator

Recap:

$$\#\text{samples} = O\left(\frac{\mathrm{poly}(H)\,|S||A|}{\varepsilon^2}\right)$$

linear in #numbers of states

What if: #states are exponentially large? (e.g. Atari games)

↓

we almost cannot visit one state twice

<u>Question:</u> How to generalize from observed states to unobserved states?

↑

similarity btw states

↑

function approx + (assumptions)

simplest model: linear MDP

Episodic, time-inhomogeneous MDP: $M = (S, A, \{r_h\}_{h \in [H]}, \{P_h\}_{h \in [H]}, H)$

$r_h : S \times A \to \mathbb{R}$, $P_h(\cdot \mid s,a) \in \Delta(S)$

<u>Linear MDP:</u> $\phi : S \times A \longrightarrow \mathbb{R}^d$ is a known feature map

$\forall h \in [H]$, $\exists w_h \in \mathbb{R}^d$: $r_h(s,a) = \langle \phi(s,a), \theta_h \rangle \ \forall (s,a)$

$\exists \mu_h \in \{S \longrightarrow \mathbb{R}^d\}$: $P_h(s' \mid s,a) = \langle \phi(s,a), \mu_h(s') \rangle$

e.g. tabular MDP: $d = |S| \cdot |A|$, $\phi(s,a) = e_{(s,a)}$

<u>Lemma:</u> Q_h^π is linear in ϕ $\forall h, \pi$

$\forall \pi$, $\exists$ $w(\pi) \in (\mathbb{R}^d)^H$ $Q_h^\pi(s,a) = \langle \phi(s,a), w_h(\pi) \rangle$

Linear MDP w/ simulator

$$(s,a,h) \longrightarrow \boxed{\text{sim}} \longrightarrow s' \sim P_h(\cdot \mid s,a)$$

assume r_h is known for simplicity

how to estimate the optimal policy?

Least-square value iteration (LSVI):

- Construct a "core set" $K \in S \times A$ of state-action pairs
- Collect data w/ the simulator:

For $h = 1 : H$ do

For each $(s,a) \in K$,

query (s,a,h) n times $\longrightarrow$ $s'_1, \ldots, s'_n \overset{\text{iid}}{\sim} P_h(\cdot \mid s,a)$

Add $\{(s,a,s'_i)\}_{i \in [n]}$ to D_h

- Backup estimation:

$\hat{V}_{H+1}(s) = 0 \quad \forall s$

for $h = H, H-1, \ldots, 1$:

$$\hat{w}_h = \underset{w \in \mathbb{R}^d}{\arg\min} \sum_{(s,a,s') \in D_h} \left(\phi(s,a)^T w - r_h(s,a) - \hat{V}_{h+1}(s') \right)^2$$

$$\hat{Q}_h = \phi^T \hat{w}_h$$

$$\hat{V}_h(s) = \max_a \hat{Q}_h(s,a)$$

$$\hat{\pi}_h(s) \in \arg\max_a \hat{Q}_h(s,a)$$

Return: $\hat{\pi} = \{\hat{\pi}_h\}_{h \in [H]}$

Assume: $\text{span}\{\phi(s,a) \mid (s,a) \in S \times A\} = \mathbb{R}^d$

core set

$$K = \{(\bar{s}_i, \bar{a}_i)\}_{i \in [d]} \qquad \text{span}\{\phi(\bar{s}_i, \bar{a}_i) \mid (\bar{s}_i, \bar{a}_i) \in K\} = \mathbb{R}^d$$

D-optimal design

$$\# \text{samples} = d \cdot n \cdot H$$

Goal: Bound $V_1^{\pi^*}(s) - V_1^{\hat{\pi}}(s)$

Define: $$\Lambda = \sum_{i \in [d]} \phi(\bar{s}_i, \bar{a}_i)\, \phi^T(\bar{s}_i, \bar{a}_i)$$

Dirac distribution

$$= \frac{1}{n} \sum_{(\bar{s}, \bar{a}) \in D_h} \phi(\bar{s}, \bar{a})\, \phi^T(\bar{s}, \bar{a})$$

$$\Psi = (\phi(\bar{s}_i, \bar{a}_i))_{i \in [d]} \in \mathbb{R}^{d \times d}$$

investigate the least-square solution:

$$\hat{w}_h = \arg\min_{w \in \mathbb{R}^d} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \left(\phi(\bar{s},\bar{a})^T w - r_h(\bar{s},\bar{a}) - \hat{V}_{h+1}(\bar{s}') \right)^2$$

$$= \frac{1}{n} \Lambda^{-1} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \phi(\bar{s},\bar{a}) \left(r_h(\bar{s},\bar{a}) + \hat{V}_{h+1}(\bar{s}') \right)$$

$$= \frac{1}{n} \Lambda^{-1} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \phi(\bar{s},\bar{a}) \left(\phi(\bar{s},\bar{a})^T \theta_h + \hat{V}_{h+1}(\bar{s}') \right)$$

$$= \theta_h + \frac{1}{n} \Lambda^{-1} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \phi(\bar{s},\bar{a}) \hat{V}_{h+1}(\bar{s}')$$

$$\phi(s,a)^T \hat{w}_h = \phi(s,a)^T \theta_h + \frac{1}{n} \phi(s,a)^T \Lambda^{-1} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \phi(\bar{s},\bar{a}) \hat{V}_{h+1}(\bar{s}')$$

$$= r_h(s,a) + \left[\hat{P}_h \hat{V}_{h+1} \right](s,a)$$

where

$$\hat{P}_h(s'|s,a) = \frac{1}{n} \phi(s,a)^T \Lambda^{-1} \sum_{(\bar{s},\bar{a},\bar{s}') \in D_h} \phi(\bar{s},\bar{a}) \, \delta_{\bar{s}'}(s')$$

$$= \phi(s,a)^T \bar{\Lambda}^{-1} \sum_{j=1}^{d} \phi(\bar{s}_j, \bar{a}_j) \frac{1}{n} \sum_{i=1}^{n} \delta_{\bar{s}'_{j,i}}(s')$$

$$[P_h V](s,a) = \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a)}[V(s')]$$

Lemma (Error decomposition):

$$V_1^*(s_1) - V_1^{\hat{\pi}}(s_1) = \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H}\left[(P_h - \hat{P}_h)\hat{V}_{h+1}\right](s_h, a_h)\right]$$

$$- \mathbb{E}_{\hat{\pi}}\left[\sum_{h=1}^{H}\left[(P_h - \hat{P}_h)\hat{V}_{h+1}\right](s_h, a_h)\right]$$

$$+ \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H} \underbrace{\langle \hat{Q}_h(s_h, \cdot), \pi^*(\cdot|s_h) - \hat{\pi}(\cdot|s_h)\rangle_A}_{\leq 0}\right]$$

We only need to bound:

$$\left[(P_h - \hat{P}_h)\hat{V}_{h+1}\right](s,a)$$

$$
[(P_h - \hat{P}_h)\hat{V}_{h+1}](s,a) = \phi(s,a)^T \sum_{s' \in S} \mu_h(s') \hat{V}_{h+1}(s') - \phi(s,a)^T \Lambda^{-1} \sum_{j=1}^{d} \phi(\bar{s}_j, \bar{a}_j) \frac{1}{n} \sum_{i=1}^{n} \hat{V}_{h+1}(\bar{s}_j^i)
$$

$$
= \phi(s,a)^T \Lambda^{-1} \left[\sum_{j=1}^{d} \phi(\bar{s}_j, \bar{a}_j) \phi^T(\bar{s}_j, \hat{a}_j) \sum_{s' \in S} \mu_h(s') \hat{V}_{h+1}(s') - \sum_{j=1}^{d} \phi(\bar{s}_j, \hat{a}_j) \frac{1}{n} \sum_{i=1}^{n} \hat{V}_{h+1}(\bar{s}'_{j,i}) \right]
$$

$$
= \phi(s,a)^T \Lambda^{-1} \left[\sum_{j=1}^{d} \phi(\bar{s}_j, \hat{a}_j) \underbrace{\left[\mathbb{E}_{s' \sim P_h(\cdot \mid \bar{s}_j, \bar{a}_j)} [\hat{V}_{h+1}(s')] - \frac{1}{n} \sum_{i=1}^{n} \hat{V}_{h+1}(\bar{s}'_{j,i}) \right]}_{H \sqrt{\frac{\log(Hd/\delta)}{n}}} \right]
$$

$$
= \phi(s,a)^T (\Psi \Psi^T)^{-1} \Psi \varepsilon
$$

$$
= \phi(s,a)^T (\Psi^T)^{-1} \Psi^{-1} \Psi \varepsilon
$$

$$= \phi(s,a)^T (\psi^T)^{-1} \varepsilon$$

$$\leq \left\| \phi(s,a)^T (\psi^T)^{-1} \right\|_1 \cdot \|\varepsilon\|_\infty$$

$$\lesssim L \cdot H \sqrt{\frac{\log(H d / \delta)}{n}}$$