08/03/23

Lower bounds for OPE w/ linear realizability and uniform coverage.

Recap

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

$$r_h(s,a) = \phi(s,a)^T \theta_h, \quad P_h(s'|s,a) = \langle \phi(s,a), \mu_h(s') \rangle$$

- D-optimal design covariance :

$$\Sigma = \mathbb{E}_{(s,a)\sim\rho}\left[\phi(s,a)\,\phi(s,a)^T\right]$$

where $\rho \in \underset{\upsilon \in \Delta(X\times A)}{\arg\max} \log\det\left(\mathbb{E}_{(s,a)\sim \upsilon}\left[\phi(s,a)\cdot\phi(s,a)^T\right]\right)$

$$|\operatorname{supp}(\rho)| \leq \frac{d(d+1)}{2}$$

property: $\forall x \in \mathbb{R}^d, \quad x^T \Sigma^{-1} x \leq d$

- Offline RL from exploratory datasets

$$H \text{ datasets} \quad D_h = \{ (s_i, a_i, s_i', r(s_i, a_i)) \}_{i=1}^{N}$$

Given i, h: independent samples $s_i' \sim P_h(\cdot \mid s_i, a_i)$

Assumption "the dataset is exploratory over all dimension"

$$\frac{1}{N} \sum_{i=1}^{N} \phi(s_i, a_i) \phi(s_i, a_i)^T \geq \frac{1}{\kappa} \Sigma$$

where Σ is the D-optimal design covariance

Alg: let $\hat{\pi} = LSVI(\{D_h\}_{h \in [H]}, \phi)$

Then:

$$V_1^{*}(s_1) - V_1^{\hat{\pi}}(s_1) \leq \varepsilon \quad \text{if } N \geq \frac{poly(H, d, \log(1/\delta), \kappa)}{\varepsilon^2}$$

w.p. $1-\delta$

<u>Today</u> what are necessary conditions for generalizations?

In particular, we will evaluate realizability in RL
(RL w/ $H=1$)

In supervised learning, realizability is sufficient

e.g. PAC w/ 0-1 loss: $\text{risk}(h_{ERM}) = \Theta\left(\frac{d_{VC}}{n}\right)$
w/ realizability

informally, in RL, only realizability is not sufficient for existence of an sample-efficient algorithm, in the information-theoretic (minimax) sense.

• <u>Offline policy evaluation (OPE) problem</u> (perhaps the simplest problem in all RL problems)

Given $\pi: S \longrightarrow \Delta(A)$ and a feature mapping $\phi: S \times A \longrightarrow \mathbb{R}^d$, the goal:

output an accurate estimate of V^{π} using collected datasets $\{D_h\}_{h=1}^{H}$

using as few samples as possible.

- <u>Realizability assumption</u> (R)

$$\forall \pi : S \longrightarrow \Delta(A), \quad \exists\, \theta_1^\pi, \dots, \theta_H^\pi \in \mathbb{R}^d :$$

$$Q_h^\pi(s,a) = \phi(s,a)^T \theta_h^\pi \quad \forall (h,s,a)$$

- <u>Data coverage assumption</u> (strongest possible) (D)

$$\mathbb{E}_{(s,a)\sim\mu_h}\left[\phi(s,a)\,\phi^T(s,a)\right] = \frac{1}{d} I$$

(μ_h satisfy D-optimal design)

- <u>Theorem</u> Assume (R). For any algorithm that takes as input both a policy π and a feature mapping. $\exists$ MDP satisfying (D) s.t.:

  $\forall \pi : S \to \Delta(A)$, the algorithm requires $\Omega\left(\left(\frac{d}{2}\right)^H\right)$ samples to output the value of π up to a constant additive error w/p a.l. 0.9

- Lemma (Distiguish Bernoulli random variables)
  - $\alpha \sim \mathrm{Unif}(\{\alpha^+, \alpha^-\})$ where

$$\alpha^- = \frac{1}{2} - \frac{\varepsilon}{2}$$
$$\alpha^+ = \frac{1}{2} + \frac{\varepsilon}{2}$$

  - $x_1, \dots, x_n \overset{iid}{\sim} \mathrm{Ber}(\alpha)$
  - $\forall f : \{0,1\}^n \longrightarrow \{\alpha^-, \alpha^+\}$

$$\Pr(f(x_1, \dots, x_n) \neq \alpha) > \underbrace{\frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-n\varepsilon^2}{1-\varepsilon^2}\right)}\right)}_{\delta \in (0, 0.25)}$$

$$n = \frac{1-\varepsilon^2}{\varepsilon^2} \ln\left(\frac{1}{8\delta(1-\delta)}\right)$$

If:

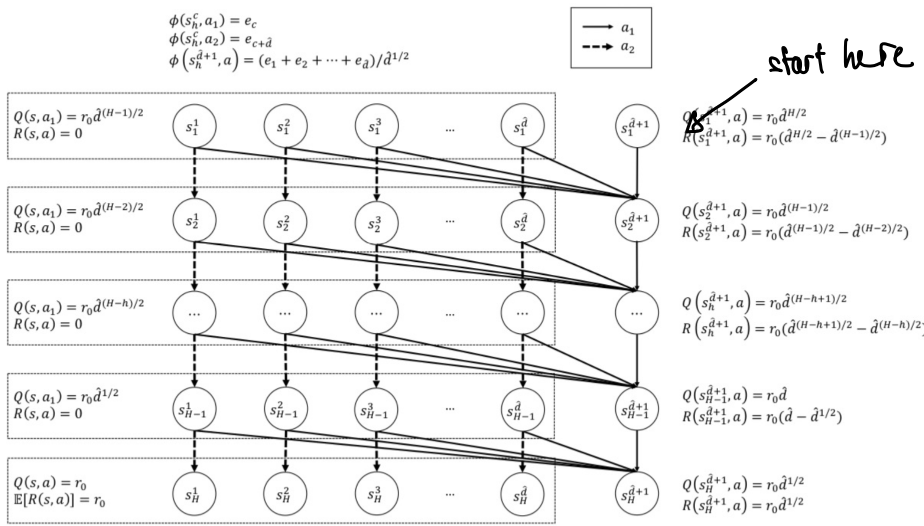
$$0.9 \leq \Pr(f(x_1, \dots, x_n) = \alpha) < 1 - \delta$$
$$\Rightarrow \delta \geq 0.1 \Rightarrow n \geq \Omega\left(\frac{1-\varepsilon^2}{\varepsilon^2}\right)$$

- Hard instances of MDP

Goal: construct MDPs that satisfy realizability (R)

construct offline data that satisfy (D)

reduction to testing problems



start here

$$Q_h^\pi(s_h^c, a_1) = r_0 \hat d^{(H-h)/2}$$

$$Q_h^\pi(s_h^c, a_2) = r_0 \hat d^{(H-h)/2}$$

$$Q_h^\pi(s_h^{\hat d+1}, a) = r_0 \hat d^{(H-h+1)/2}$$

$$Q_h^\pi(s,a) = \phi(s,a)^T w_h^\pi$$

$$R(s_H^c, a) = \begin{cases} 1 & \text{wp } \frac{(1+r_0)}{2} \\ -1 & \text{wp } \frac{1-r_0}{2} \end{cases}$$

$$Q_h^\pi(s,a) = \phi(s,a)^T \begin{bmatrix} r_0 \hat d^{(H-h)/2} \\ \vdots \\ r_0 \hat d^{(H-h)/2} \\ \vdots \end{bmatrix}$$

$\hat d$

- Dataset: μ_h is uniform over $\{(s_h^c, a_1), (s_h^c, a_2)\}_{c \in [\hat{d}]}$

$$\mathbb{E}_{(s,a)\sim \mu_h}\left[\phi(s,a)\,\phi(s,a)^T\right] = \frac{1}{\hat{d}} I$$

- Reduction to testing:

  - the policy value to estimate: $V_1^\pi(s_1^{\hat{d}+1}) = r_0 \hat{d}^{H/2}$
  - consider 2 instances: (I) $r_0 = 0 \longrightarrow V_1^\pi(s_1^{\hat{d}+1}) = 0$

    (II) $r_0 = \hat{d}^{-H/2} \longrightarrow V_1^\pi(s_1^{\hat{d}+1}) = 1$
  - if the algorithm wants to output an estimate that is correct up to 0.5 error, then it must need to distinguish two problem instances (I) and (II)

Consider any algorithm:

$$\text{Alg}: \left(\{D_h\}_{h\in[H]}, \phi\right) \longrightarrow \mathbb{R}$$

- Given datasets $\{\mathcal{D}_h\}_{h\in[H]}$ (and Φ), the algorithm need to identify which of the two instances (I) and (II) that the datasets come from

- Note that for both (I) and (II):
  - data distribution, transition kernels are the same
  - rewards are zero everywhere except in the last layer H

- Thus, to distinguish (I) and (II), the algorithm need to distinguish the reward distribution:

$$r = \begin{cases} 1 & \text{wp } \frac{1}{2} \\ -1 & \text{wp } \frac{1}{2} \end{cases}$$

and

$$r = \begin{cases} 1 & \text{wp } \frac{1+\tilde{d}^{-H/2}}{2} \\ -1 & \text{wp } \frac{1-\tilde{d}^{-H/2}}{2} \end{cases}$$

$$\Longrightarrow \quad n = \Omega\left(\hat{d}^{H}\right) = \Omega\left(\left(\frac{d}{2}\right)^{H}\right)$$