

Batch RL

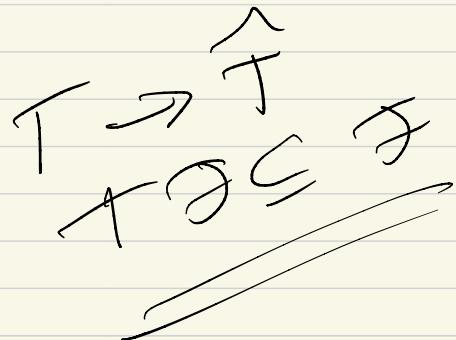
March 23

Can we adopt our "featureized"- planning algs?

Planning → Batch?



AVI
VI
LSVI



Evaluating policies

$$\pi, q: S \times A \rightarrow \mathbb{R}^d$$

L^∞ ($\| \cdot \|_\infty$ -error)
needs to
be controlled.

$$D_n = ((S_i, A_i, R_i, S'_i))_{i=1}^n \quad S'_i \sim P_{A_i}(S_i)$$

$$i = 1, \dots, n \quad \mathbb{E}[R_i | S_i, A_i] = r_{A_i}(S_i)$$

$$q^\pi(s, a) = \mathbb{E}_{s, a}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_t \right] = \int r(\tau) dP_{sa}^\pi(\tau)$$

Importance Sampling

$$\tau = (s_0, a_0, s_1, a_1, \dots)$$

$$r(\tau) \xrightarrow{\downarrow} \sum_{t=0}^{\infty} \gamma^t r_{a_t}(s_t)$$

$$\Pi_{\text{beh}} = \Pi_b$$

↓

$$dP_{\mu}^{\Pi_{\text{beh}}}(\tau)$$

$$\left\{ \begin{array}{l} S_0 \sim \underline{\mu} \in \mathcal{M}_1(S) \\ A_0 \sim \Pi_{\text{beh}}(\cdot | S_0) \\ S_1 \sim P_{A_0}(S_0) \\ \vdots \end{array} \right.$$

$$q^{\Pi}(s, a) = \int r(s) dP_{sa}^{\Pi}(\tau) = \int r(\tau) \frac{dP_{sa}^{\Pi}}{dP_{\mu}^{\Pi_{\text{beh}}}}(\tau).$$

$$\tau_1, \tau_2, \dots, \tau_n \sim P_{\mu}^{\Pi_{\text{beh}}} \quad \text{i.i.d.}$$

$$q^{\Pi}(s, a) \approx \frac{1}{n} \sum_{t=1}^n r(\tau_i) \boxed{\frac{dP_{sa}^{\Pi}}{dP_{\mu}^{\Pi_{\text{beh}}}}(\tau_i)}$$

Cut traj:

$$\tau = (s_0, a_0, \dots, s_{n-1}, a_{n-1})$$

$$P_{\mu}^{\Pi_{\text{beh}}}(\tau) = \mu(s_0) \Pi_{\text{beh}}(a_0 | s_0) \underbrace{P_{a_0}(s_0, s_1)}_{\text{red}} \Pi_{\text{beh}}(a_1 | s_1) \dots \Pi_{\text{beh}}(a_{n-1} | s_{n-1})$$

$$P_{sa}^{\Pi}(\tau) = \mathbb{I}(s_0=s) \mathbb{I}(a_0=a) \underbrace{P_{a_0}(s_0, s_1)}_{\text{red}} \Pi(a_1 | s_1) \dots \Pi(a_{n-1} | s_{n-1})$$

$$W(\pi) = \frac{d P_{sa}^{\pi}}{d P_{\pi_{\text{beh}}}^{\pi_{\text{beh}}}} (\pi) = \frac{\prod_{i=1}^{t-1} \frac{\pi(a_i | s_i)}{\pi_{\text{beh}}(a_i | s_i)}}{\prod_{i=0}^{t-1} M(s_i)^T \pi_{\text{beh}}(a_0 | s_0)}$$

We can compute this!

π / π_{beh} not aligned

$\Rightarrow W(\pi)$ very small most of the time.

Garbage in - garbage out

Rare - event

I.S.

\rightarrow W.I.S. / S.N.I.S.

$$\frac{\sum w(s_i) r(\pi_i)}{\sum w(s_i)}$$

biased variance

I.S.

create targets + least-squares regression value-targets.

$$\max_{z \in S \times A} \|\varphi(z)\|_{G_S^{-1}}$$

$$G_S = \sum_{z \in C} \mu(z) \varphi(z) \varphi(z)^T$$

$$z_1, \dots, z_n \sim \underline{\mu} \in \mathcal{M}_1(S \times A)$$

$$\hat{G} = \frac{1}{n} \sum_{i=1}^n \varphi(z_i) \varphi(z_i)^T ; \lambda_{\min}(\hat{G}) > 0$$

$$\max_{z \in S \times A} \|\varphi(z)\|_{G_S^{-1}}^2 \leq \max_{z \in S \times A} \|\varphi(z)\|_2^2 \overline{\lambda_{\min}(G)}$$

$$x^T A x \leq \lambda_{\max}(A) \|x\|_2^2$$

$A \succeq 0$

$$\lambda_{\max}(G^{-1}) = \lambda_{\min}(G)$$

$$\lambda_{\min} \left(\sum_{z \in Z} \mu(z) \varphi(z) \varphi(z)^T \right) > 0$$

$$\text{TD: } q^\pi = T_\pi q^\pi$$

$$q^\pi = \Phi \theta \quad \Phi \theta = T_\pi \Phi \theta \quad , \quad \theta \in \mathbb{R}^d$$

\downarrow
SA constraints

overconstrained!

$$\begin{aligned} \varphi(s_i, a_i)^T \theta &= r_{a_i}(s_i) + \gamma \underbrace{\frac{P_{a_i}(s_i)}{M_\pi} \Phi \theta}_{\text{M}_\pi \Phi \theta} \\ &\approx R_i + \gamma \underbrace{(M_\pi \Phi \theta)(s_i)}_{R_i} \\ &= R_i + \gamma \underbrace{\varphi(s'_i, \pi(s'_i))^T \theta}_{\varphi_\pi} \end{aligned}$$

$$\delta_i(\theta) = \varphi(s_i, a_i)^T \theta - (R_i + \gamma \varphi(s'_i, \pi(s'_i))^T \theta)$$

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \delta_i^2(\theta) \rightarrow \min \quad \text{bias}$$

Problem: No consistency

$$q^\pi = \Phi \theta$$

$$\underbrace{\frac{1}{n} \sum_{i=1}^n \delta_i(\theta) \varphi(s_i, a_i)}_{\text{LSTD}} = 0 \quad \text{Least-squares}$$

$$q^\pi = \underbrace{T_n q}_{} \quad \boxed{q_{i+1} = \hat{T}_\pi q_i}$$

$$T_\pi \mathcal{F} \subseteq \mathcal{F} \quad (\text{lin. MDPs})$$

$$D) \left\| \frac{dP_{sa}^{\pi}}{dP_{\mu}^{\text{beh}}} \right\|_{\infty} < \infty$$

2. ~~$T_x \mathcal{F} \subseteq \mathcal{F}$~~

$T \mathcal{F} \subseteq \mathcal{F}$

~~$\lambda_{\min}(E_{\mu}(\mathbf{U}\mathbf{U}^T)) \geq c > 0$~~

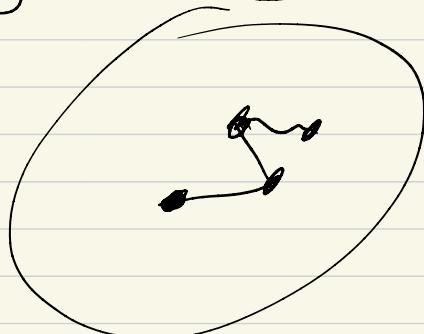
i.i.d

[Markov ass.]

Z

mixing

Bernstein



forgetting
the past /
decoupling
per part

