Batch RL

March 23

Can we adopt our "featurized" planning algos?

Planning → Batch?

- API
- VI
- FQI
- LSPI
- ⋮
- Politex

AVI
VI
LSVI

$T \to \hat{T}$

$T \mathcal{J} \subseteq \mathcal{J}$

Evaluating policies

π , $q : S \times A \to \mathbb{R}^d$

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

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

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

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

Importance Sampling

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

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

$\pi_{beh} = \pi_b$

$\downarrow$

$dP_\mu^{\pi_{beh}}(\tau)$

$$\begin{cases} S_0 \sim \mu \in \mathcal{M}_1(S) \\ A_0 \sim \pi_{beh}(\cdot \mid S_0) \\ S_1 \sim P_{A_0}(S_0) \\ \vdots \end{cases}$$

$$q^\pi(s,a) = \int r(\tau)\, dP_{sa}^\pi(\tau) = \int r(\tau) \frac{dP_{sa}^\pi}{dP_\mu^{\pi_{beh}}}(\tau)\, dP_\mu^{\pi_{beh}}(\tau)$$

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

$$q^\pi(s,a) \approx \frac{1}{n} \sum_{i=1}^{n} r(\tau_i) \frac{dP_{sa}^\pi}{dP_\mu^{\pi_{beh}}}(\tau_i) = w(\tau_i)$$

Cut traj:

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

$$P_\mu^{\pi_{beh}}(\tau) = \mu(s_0)\, \pi_{beh}(a_0 \mid s_0)\, P_{a_0}(s_0, s_1)\, \pi_{beh}(a_1 \mid s_1) \cdots \pi_{beh}(a_{n-1} \mid s_{n-1})$$

$$P_{sa}^\pi(\tau) = \mathbb{I}(s_0 = s)\, \mathbb{I}(a_0 = a)\, P_{a_0}(s_0, s_1)\, \pi(a_1 \mid s_1) \cdots \pi(a_{n-1} \mid s_{n-1})$$

$$W(\tau) = \frac{dP^{\pi}_{s_0 a_0}}{dP^{\pi_{beh}}_{n}}(\tau) = \frac{\mathbb{I}(s=s_0, a=a_0)}{\mu(s_0)\pi_{beh}(a_0|s_0)} \prod_{i=1}^{h-1} \frac{\pi(a_i|s_i)}{\pi_{beh}(a_i|s_i)}$$

We can compute this!

π / π_{beh} not aligned

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

Garbage in - garbage out

Rare-event

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

$$\frac{\sum W(\tau_i) R(\tau_i)}{\sum W(\tau_i)}$$

biased

variance

I.S. create targets + least-square regression value-targets.

$$\max_{z \in S \times A} \|\varphi(z)\|_{G_\rho^{-1}} \qquad G_\rho = \sum_{z \in C} \rho(z)\varphi(z)\varphi(z)^T$$

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

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

$$\max_{z \in S \times A} \|\varphi(z)\|^2_{G^{-1}} \le \max_{z \in S \times A} \|\varphi(z)\|_2^2 \, \lambda_{\min}^{-1}(\hat{G})$$

$$x^T A x \le \lambda_{\max}(A)\|x\|_2^2 \qquad A \succeq 0$$

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

G_μ

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

TD: $\quad q^\pi = T_\pi q^\pi$

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

↑ SA constraints

Unconstrained!

$$\varphi(s_i,a_i)^T\theta = r_{a_i}(s_i) + \gamma P_{a_i}(s_i) M_\pi \Phi\theta$$

$$\approx \underbrace{r_{a_i}(s_i)}_{R_i} + \gamma (M_\pi \Phi\theta)(s_i')$$

$$= R_i + \gamma \varphi(s_i', \pi(s_i'))^T\theta$$

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

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

Problem: No consistency

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

$$\frac{1}{n}\sum_{i=1}^n \delta_i(\theta)\varphi(s_i,a_i) = 0$$

LSTD

Least-squares

$$q^\pi = T_\pi q$$

$$q_{i+1} = \hat{T}_\pi q_i$$

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

1) $\left\| \frac{dP^{\pi}_{sa}}{dP^{\pi_{beh}}_{\mu}} \right\|_{\infty} < \infty$

2. $T_{\pi} \mathcal{F} \subseteq \mathcal{F}$

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

$\lambda_{min}\left(\mathbb{E}_{\mu} \varphi \varphi^{T}\right) \geq c > 0$

---

i.i.d

Markov ass.

↓

mixing

↓

Bernstein

forgetting the past / decoupling per past

Z