Batch RL ≈ fake RL → no interaction!

Data! → Policies → MDP → ~~Data!~~

Control

Who collected → How was the collected?

Do we know this?

Yes / No

Experimental Design: How to collect data

How to use the data? O.P.O.

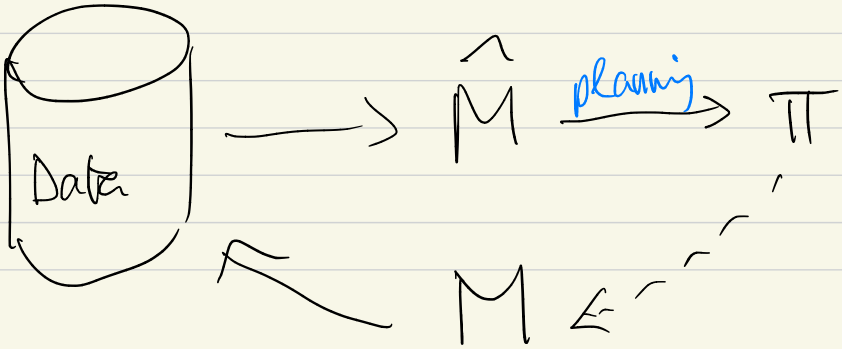
↓ O.P.E

Certification of performance

conservative exploration

Limits / how to approach?

Model-based P.O.



Model-sensitivity

$0 \le \gamma < 1$

$\hat M = (S, A, \hat P, \hat r, \gamma)$ $\qquad$ $M = (S, A, P, r, \gamma)$

$\pi(a|s)$

$$\boxed{(*) \quad \hat v^\pi \ge \hat v^* - \varepsilon \mathbb{1}} \implies v^\pi \ge v^* - \delta \mathbb{1}$$

$$\delta \le f\left(\varepsilon, \frac{1}{1-\gamma}, M, \hat M\right)$$

$$f = ?$$

Policy error bound:

1. $M_\pi q^* \ge v^* - \varepsilon \mathbb{1} \implies v^\pi \ge v^* - \frac{\varepsilon}{1-\gamma} \mathbb{1}$

2. $M_\pi q = M q \implies v^\pi \ge v^* - \frac{2\|q - q^*\|_\infty}{1-\gamma} \mathbb{1}$

$q = \hat q^\pi \qquad \hat q^\pi \le \hat q^* \implies M \hat q^\pi \le M \hat q^* = \hat v^*$

$$M \hat q^\pi \ge M_\pi \hat q^\pi = \hat v^\pi \overset{(*)}{\ge} \hat v^* - \varepsilon \mathbb{1} = M \hat q^* - \varepsilon \mathbb{1} \ge M \hat q^\pi - \varepsilon \mathbb{1}$$

$$M_\pi \hat q^\pi = M \hat q^\pi + z, \quad \|z\|_\infty \le \varepsilon$$

$$M_\pi q = M q + z$$

$$M_\pi q^* = M_\pi q + M_\pi (q^* - q) = M q + M_\pi (q^* - q) + z$$

$$= Mq^* + \underbrace{Mq - Mq^*} + M_\pi(q^* - q) + z$$

$$\|Mq - Mq^*\|_\infty \le \|q - q^*\|_\infty$$

$$Mq - Mq^* \ge -\|q - q^*\|_\infty \mathbb{1}$$

$$\|M_\pi(q^* - q)\|_\infty \le \|q - q^*\|_\infty \Rightarrow M_\pi(q^* - q) \ge -\|q - q^*\|_\infty \mathbb{1}$$

$$z \ge -\|z\|_\infty \mathbb{1}$$

$$M_\pi q \ge Mq - (\underbrace{2\|q - q^*\|_\infty + \|z\|_\infty}_{c}) \mathbb{1}$$

$$\Rightarrow \boxed{v^\pi \ge v^* - \frac{2\|q - q^*\|_\infty + \|z\|_\infty}{1-\gamma} \mathbb{1}}$$

$$q := \hat{q}^\pi$$

$$v^\pi \ge v^* - \boxed{\frac{2\|\hat{q}^\pi - q^*\|_\infty + \varepsilon}{1-\gamma}} \mathbb{1}$$

$$\|\hat{q}^\pi - q^*\|_\infty \le \underbrace{\|\hat{q}^\pi - \hat{q}^*\|_\infty}_{\le \varepsilon} + \underbrace{\|\hat{q}^* - q^*\|_\infty}_{?}$$

general:

T γ-contraction, $Tx = x$

$\forall y \quad \|x - y\| \leq \frac{\|Ty - y\|}{1-\gamma}$

$\|x - y\| \leq \|Tx - Ty\| + \|Ty - y\| \leq \gamma\|x - y\| + \|Ty - y\|$

$x = \hat{q}^* \qquad T := \hat{T}$

$\hat{T}q = \hat{r} + \gamma\hat{P}Mq$

$Tq = r + \gamma PMq$

$\|\hat{q}^* - q^*\|_\infty \leq \frac{\|\hat{T}q^* - q^*\|_\infty}{1-\gamma}$

$\|\hat{T}q^* - q^*\|_\infty = \|\hat{T}q^* - Tq^*\|_\infty$

$\leq \|\hat{r} + \gamma\hat{P}v^* - (r + \gamma P v^*)\|_\infty$

$\leq \|r - \hat{r}\|_\infty + \gamma\|(\hat{P} - P)v^*\|_\infty$

$\leq \|\hat{P} - P\|_\infty \|v^*\|_\infty$

$\leq \frac{1}{1-\gamma}$

$\|\hat{q}^* - q^*\| \leq \frac{\|T\hat{q}^* - \hat{q}^*\|_\infty}{1-\gamma}$

$\|T\hat{q}^* - \hat{q}^*\|_\infty \leq \|r - \hat{r}\|_\infty + \gamma\|(\hat{P} - P)\hat{v}^*\|_\infty$

Thm : $\forall \pi \quad \hat{v}^{\pi} \geq \hat{v}^{*} - \varepsilon \mathbb{1}$

$$\Rightarrow \quad v^{\pi} \geq v^{*} - \delta \mathbb{1}$$

$\min(a,b) = a \wedge b$

$$\delta \leq \frac{\varepsilon \cancel{(1+\gamma)}^{\,2}}{1-\gamma} + \frac{2}{(1-\gamma)^2}\Big[\big(\|r-\hat{r}\|_\infty + \gamma \underbrace{\|(\hat{P}-P)v^{*}\|_\infty}\big) \wedge \big(\|r-\hat{r}\|_\infty + \gamma\|(\hat{P}-P)\hat{v}^{*}\|_\infty\big)\Big]$$

$$\leq \frac{\varepsilon \cancel{(1+\gamma)}^{\,2}}{1-\gamma} + \frac{2}{(1-\gamma)^2}\left[\|r-\hat{r}\|_\infty + \frac{\gamma\|P-\hat{P}\|_\infty}{1-\gamma}\right]$$

$r \in [0,1]$

Batch RL

1. **Tabular**   2. **Featurized**

1. trajectories; π_b 2. offline access to sim $n(s,a)$

behavior policy
logging policy

$t = 1, \dots, n$

(S_t, A_t, S'_t, R_t)

$S'_t \sim P_{A_t}(S_t, \cdot)$

states

$R_t \sim \tilde{P}_{A_t}(S_t, \cdot)$

reals

$$\hat{r}_a(s) = \begin{cases} \dfrac{1}{N(s,a)} \displaystyle\sum_{t=1}^{n} \mathbb{I}(S_t = s, A_t = a) R_t \\ 0, & N(s,a) = 0 \end{cases}$$

$$N(s,a) = \sum_{t=1}^{n} \mathbb{I}(S_t = s, A_t = a)$$

$N(s,a) > 0.$

$$\hat{P}_a(s,s') = \begin{cases} \dfrac{1}{N(s,a)} \displaystyle\sum_{t=1}^{n} \mathbb{I}(S_t = s, A_t = a)\,\mathbb{I}(S'_t = s') \\ 1/|S|, & N(s,a) = 0 \end{cases}$$

$\underline{N(s,a) = n(s,a)}, \quad > 0 \qquad R_t \in [0,1]$

$$\left.\begin{aligned} |\hat{r}_a(s) - r_a(s)| &\le \sqrt{\frac{\log\left(\frac{SA}{\delta}\right)}{2n(s,a)}} \\ \underline{|\hat{P}_a(s,s') - P_a(s,s')|} &\le \sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{2n(s,a)}} \end{aligned}\right\} \text{ wp } 1-\delta \quad \forall s,a,s'$$

$$\boxed{\begin{aligned} \|\hat{P} - P\|_\infty &= \max_{s,a} \underbrace{\|\hat{P}_a(s) - P_a(s)\|_1} \\ &\le S \sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{2n(s,a)}} \\ \|r - \hat{r}\|_\infty &\le \sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{2n(s,a)}} \le \sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{n_{min}}} \end{aligned}}$$

$n_{min} = \min_{(s,a)} n(s,a)$

$$\delta \le \frac{2}{(1-\gamma)^2}\left[\sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{2n_{min}}} + \frac{S}{1-\gamma}\sqrt{\frac{\log\left(\frac{S^2 A}{\delta}\right)}{2n_{min}}}\right] = UB(n)$$

$\Downarrow$ max n_{min}

$n(s,a) = ?$

$$\boxed{\sum_{s,a} n(s,a) = n}$$

$$n(s,a) = \frac{n}{SA} = n_{min}$$

$$\underbrace{\delta \le \frac{2}{1-\gamma^2}\left[\sqrt{\frac{SA \log}{2n}}\left(1+\frac{S}{1-\gamma}\right)\right]}$$

$$\varepsilon = H^3\left[\sqrt{\frac{SA}{n}}\, S\right]$$

$$\frac{\varepsilon}{SH^3} = \sqrt{\frac{SA}{n}}$$

$$\left(\frac{SH^3}{\varepsilon}\right)^2 SA$$

$$\boxed{n \overset{\text{log}}{\approx} SA\left(\frac{SH^3}{\varepsilon}\right)^2}$$

$$= \frac{S^3 A}{\varepsilon^2} H^6$$

trajectories $\rightsquigarrow \varphi$