Offline learning

Infinite-horizon $M = (S, A, P, R, \gamma, \cancel{d_0})$ $d_0 \in \mathcal{M}_1(S)$ initial state dist. $\xrightarrow{s_0}$ → lower bound only.

$V_M^\pi(s) = \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h \mid S_0 = s\right]$; $Q_M^\pi(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h=0}^{\infty} \gamma^h r_h \mid S_0 = s, A_0 = a\right]$

$\exists$ determ. $\not$ & ML policy π_M^* s.t. $V_M^* = V_M^{\pi_M^*} = \max_\pi V_M^\pi$, $Q_M^* = Q_M^{\pi_M^*}$, $\pi_M^* = \operatorname{argmax} Q_M^*$

Dataset ~~oracle~~ oracle-generated $D = \{(\cancel{S_i, A_i, r_i, S_i'}, S_i, A_i, R_i, S_i')\}_{i=1}^n$ $(S_i, A_i) \overset{i.i.d.}{\sim} \mu \in \mathcal{M}_1(S \times A)$

policy-induced $(\pi_{\log}, \vec{n})$ $D = \{\tau_j\}_{j=1}^m$ $\tau_j = (S_0^{(j)}, A_0^{(j)}, R_0^{(j)}, \cdots, S_{n_j-1}^{(j)}, A_{n_j-1}^{(j)}, R_{n_j}^{(j)}, S_{n_j}^{(j)})$

Goal [Data is collected before any interactions with the env.]

Batch policy opt. $(s_0, \{M \in \mathcal{M}\})$ share the same state, action & γ

when deciding the collection dist., we cannot dist.

w.p. $\geq 1-\delta$, for every instance (s_0, M),

the learnt policy $\hat{\pi}_D$ is near opt. $\mathbb{P}(|V^*(s_0) - V^{\hat{\pi}_D}(s_0)| \leq \varepsilon) \geq 1-\delta$

$\forall \mu$, $\min_w \mu(C_\delta(w) \cup C_\delta(-w)) \leq \frac{2\mathrm{Vol}(C_\delta(w))}{\mathrm{Vol}(\mathcal{B})} = \frac{2}{N(\delta}$

$\mathbb{P}_w(\{a_1, \cdots, a_n\} \not\subseteq C_\delta(w) \cup C_\delta(-w))$

$\geq \left(1 - \frac{2}{N(\delta,d)}\right)^n \geq \left(1 - \frac{2}{N(\delta,d)}\right)^{\frac{N}{2}-1} \geq \frac{1}{e}$

if $n < \frac{\cancel{e}}{N(\delta,d)/2}$

Assumptions:

1. realizability (remove dependence from $|S|$ & $|A|$)

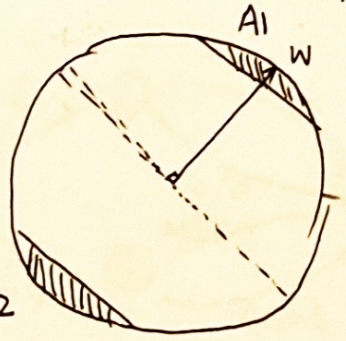
Given an instance $(s_0, \mathcal{M})$, $\exists$ feat. map $\phi(s,a) \in \mathbb{R}^d$ with $\|\phi(s,a)\|_2 \le 1$

such that $\forall M \in \mathcal{M}$ and $\forall \pi$, $\forall (s,a)$, $Q_M^\pi(s,a) = \phi(s,a)^T \theta_M^\pi$ with $\|\theta_M^\pi\|_2 \le 1$

Thm 3: $\exists$ a BPO $(s_0, \mathcal{M})$ with all-policy realizability s.t. any algo. using less than $\frac{N(\delta,d)}{2} \gtrsim \Omega\frac{1}{2}(\frac{1}{1-\gamma})^d$ samples has $V^*(s_0) - V^{\hat{\pi}_D}(s_0) \ge 1$ w.p. at least $1 - \frac{1}{2e}$ (Zanette 2020)

B
① Rewards are not seen.
② Bootstrapping issue, or the rank deficient features cause the hardness.

A Proof steps: 1. realizability holds 2. the err. of not choosing the opt. action is large 3. $\exists$ two MDPs which cannot be distinguished under n samples but a set of has diff. opt. actions



D
① rewards only exist in small areas $\{x: x^T w \ge \gamma\}$ $\le -\gamma\}$
② unshadowed area: $(\phi(s) - \gamma\phi(s'))^T w = 0$
⇒ there are multiple possible soln. without exploring the small caps
③ $r_w = 0$ in the unshadowed area.
$+w, -w$ can be soln.
M_+ M_-
~~positive rewards if $a^T w > 0$~~
~~neg. o.w.~~ For M_+, opt. arm $+w$ with $q_+(s_0, w) = 1$
$q_-(s_0, w) = -1$

C $\mathcal{B} = \{x: \|x\|_2 \le 1\}$
$\le \frac{2}{(\frac{1}{1-\gamma})^d}$ $S = \{s_0\} \cup \mathcal{B}$

$A(s) = \begin{cases} \mathcal{B}, & \text{if } s = s_0 \\ 8, & \text{o.w.} \end{cases}$

~~$\phi(s,a) = a$~~ $\phi(s,a) = a$

$s'(a) = \begin{cases} \frac{1}{8}(a^T w)\cdot w, & \text{if } a \notin G(w) \cup G(-w) \\ a, & \text{if } a \in G(w) \cup G(-w) \end{cases}$

2. Concentrability (a) feature coverage

$\forall (s,a)$, $\|\varphi(s,a)\|_2 \le 1$ and $\mathbb{E}_{(s,a)\sim\mu}[\varphi(s,a)\varphi(s,a)^T]$ has its smallest eigenval. equaling to $\frac{1}{d}$.

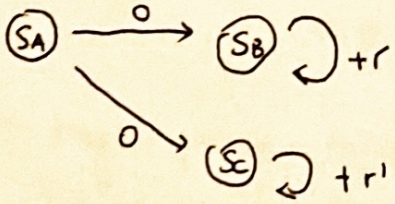
how feat. cover all the dimensions

largest var. dir. ⇒ eig. vec.

its magnitude ⇒ eig. val.

supervised learning

(Amortila 2020)



$\varphi(s_A, a_1) = [0, \gamma]^T$, $\varphi(s_C, a) = [0, 1]^T$

$\varphi(s_A, a) = [\gamma, 0]^T$

$\varphi(s_B) = [1, 0]^T$

$\mathcal{D} = \{(s_A, 0, s_B)\}$

$\theta^* = \frac{r}{1-\gamma}$

$\mathbb{E}_{s\sim M}[\phi(s)\phi(s)^T] = \gamma^2 I$

q^π realizability but q^π is not observed ; only $r + \gamma \hat{v}(s')$ is observed

[linear Bellman ~~realizability~~ closedness]

(b) concentrability : $\nu \in \mathcal{M}_1(S\times A)$ admissable, if $\exists \pi$ & h s.t. $\mathbb{P}^\pi(S_h = s, A_h = a) = \nu(s,a)$

$\exists$ a const. $C_{conc} < \infty$ s.t. $\forall$ admissable ν, $\sup_{(s,a)} \frac{\nu(s,a)}{\mu(s,a)} \le C$

$C \ge 64$

Thm : $\forall S \ge 9$ and $\gamma \in (\frac{1}{2}, 1)$, $\exists \mathcal{M}$ with $|\mathcal{S}| = S$, $|\mathcal{A}| = 2$, $|\mathcal{F}| = 2$ and a sampling dist. μ which is admissable

1. q^π realizable & $C_{conc} \le$ C

~~2. $\varepsilon \le S^{1/3}$ gives $V^{\pi^*}(s_0)$ -~~

2. $n \le c \cdot \min\{S^{1/3}(\log S)^2, 2^{C/32}, 2^{\frac{1}{1-\gamma}}\}$ must have $J(\pi^*) - \mathbb{E}_{D,M}[J(\hat{\pi})] \ge C'$

Concentration bound + all-policy realizability $\not\Rightarrow$ solve batch RL in poly $(\frac{1}{\varepsilon}, \frac{1}{1-\gamma}, C_{conc}, \log$

(Foster et al. 2021)

Sufficient Conditions (Some results hold for episodic / finite-horizon MDP)

1. concentration bound + Bellman completeness / closedness

Given a finite func. class $\mathcal{F}$, if $f \in \mathcal{F}$, $\mathcal{T}f \in \mathcal{F}$

Thm. (Chen & Jiang 2019): $\sup_{\text{admissable } \nu} \left\| \frac{\nu}{\mu} \right\|_\infty \le C < \infty$ & completeness, w.p. $1-\delta$, the output policy of FQI after k iter., $\hat{\pi}_k$ satisfies $\lim_{k\to\infty} (v^* - v^{\hat{\pi}_k}) \le \varepsilon \cdot V_{\max}$

when $n = O\left(\frac{C \ln \frac{|\mathcal{F}|}{\delta}}{\varepsilon^2 (1-\gamma)^4} \right)$ discounted

(mixmax-optimal ?)

Why not completeness? ~~stronger than realizability~~

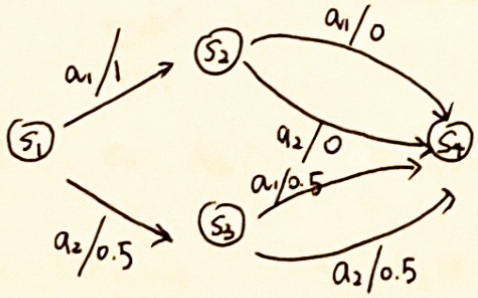
not monotone in $\mathcal{F}$, adding more func. may break the condition

(Tkachuk et al. 2024)

2. concentration bound + q^π-realizability + trajectory data + skipper

[Q: how to get a contractive learner $\hat{\mathcal{T}}$ / $\hat{\mathcal{T}}^\pi$]

target: $r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \begin{cases} \max_{a'} \phi(s',a')^\top \theta_{h+1} \\ q^\pi(s',a') \text{ where } \pi = \operatorname{argmax}_{a'} \phi(s',a')^\top \theta_{h+1} \end{cases} + \eta_h(s,a)$ → noise amplified exp.



$\phi(s_1, a_1) = \phi(s_1, a_2) = 1$

$\phi(s_2, a_1) = \phi(s_2, a_2) = 0$

$\phi(s_3, a_1) = \phi(s_3, a_2) = 0$

$q(s,a)$-realizable

but $r(s_1, a_1)$ & $r(s_1, a_2)$ is not

$\sup_{s,a} |r(s,a) - \phi(s,a)^\top \theta| \ge 0.5/2$

~~range(s) =~~

skip states with approx. $\sup_a |V^\pi(s) - q^\pi(s,a)|$ small

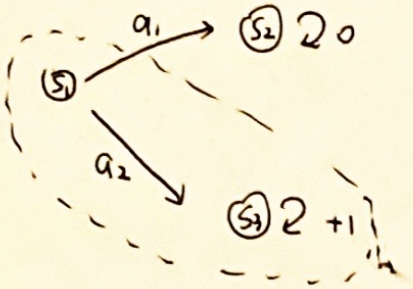
use multi-step rewards $\Rightarrow$ ~~$\forall \pi$~~ $\mathbb{E}_{\text{traj} \sim \pi, \text{skip-termination } \tau} [R_{h:\tau-1} + f(s_\tau)]$

lin. realizable $\forall f : S \to [0, H]$

stronger than optimality-closedness

$n = \tilde{\Theta}\left(C_{\text{conc}}^4 H^7 d^4 / \varepsilon^2 \right)$

3. all-policy-coverage $\Rightarrow$ single-policy-coverage



$\phi(s_1) = [0, \frac{\gamma}{1-\gamma}, 0]$ $\quad \theta^* = [1, 1, 0]$

$\phi(s_2) = [0, 0, 2]$ $\quad \hat{\theta} = [1, 1, 1]$ initialization

$\phi(s_3) = [\frac{1}{1-\gamma}, 0, 0]$

ensure no over-estimated extrapolation $\Rightarrow$ instability ← pessimism

{ penalized val. func. of under-covered policies ✓
penalize policies that are far from the one generating data

Vlad's proof $\forall \pi$, $\mathbb{E}_{(S_h, A_h) \sim P_h^\pi(\cdot,\cdot)} [\, \| \phi(S_h, A_h) \|_{X_h^{-1}}] \le \mathbb{E}_{U_h} [\, \| \phi \|_{X_h^{-1}}] \cdot C_{conc} \le \tilde{O}\left(\frac{Cd}{\sqrt{n}}\right)$

$X_h \overset{n_h}{\sum_{i=1}} \phi(S_h^i, a_h^i) \phi(S_h^i, a_h^i)^T + \lambda I$

$\Rightarrow$ single-policy-concentration $\quad \Leftrightarrow \quad \frac{d_h^\pi(S_h, a_h)}{d_h^\mu(S_h, a_h)} \le C \quad \forall h, (S_h, a_h)$ for π

[finite-horizon]

(Jin et al. 2021) ② $\quad \sum_{h=1}^{H} \mathbb{E}_\pi [\, \| \phi(S_h, a_h) \|_{X_h^{-1}}] \le C(\pi, \mathcal{D})$

(discounted) ③ $\quad \max_{f \in \mathcal{F}} \frac{\| f - \mathcal{T}^\pi f \|^2 \, \mathbb{E}_{(s,a) \sim d^\pi} [(f(s,a) - \mathcal{T}^\pi f(s,a))^2]}{\mathbb{E}_{(s,a) \sim \mu} [(f(s,a) - \mathcal{T}^\pi f(s,a))^2]}$

$=: C(\pi, \mu, \mathcal{F}) \le \| \frac{d^\pi}{\mu} \|_\infty$

Jin et al. (2021) $\quad$ single-policy + linear MDP

like UCB $\quad \hat{w}$ LSVI

$\Gamma_h(s,a) = \beta \| \phi(s,a) \|_{\Lambda_h^{-1}}$ $\quad$ not { lin. realizable complete

$\hat{Q}_h(s,a) = \min \{ \phi(s,a)^T \hat{w} - \Gamma_h(s,a), H - h + 1 \}^+$

Xie et al. (2021) { Zanette (2021) $\quad$ single-policy + $\mathcal{T}^\pi$-Bellman restricted completeness

$\forall \pi_{h+1} \in \Pi_{h+1}$, $\forall Q_{h+1} \in \mathcal{F}_{h+1}$

$\inf_{Q_h \in \mathcal{F}_h} \| Q_h - \mathcal{T}_b^{\pi_{h+1}} Q_{h+1} \|_\infty = 0$

Xie: $\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}} \underset{f \in \mathcal{F}_{\pi,\varepsilon}}{\min} f(s_0, \pi)$ s.t. $\mathcal{F}_{\pi,\varepsilon} \supseteq \{f \in \mathcal{F} :$ empirical Bellman err. under $\pi \leq \varepsilon\}$

Zanette: (Actor Critic) For $t = 1, \cdots, T$,

$$(\theta_t, W_t) = \underset{\theta \in (\mathbb{R}^d)^H, W \in (\mathbb{R}^d)^H}{\operatorname{argmin}} \sum_a \pi_{\theta_t}(a|s_1) \langle \phi_1(s,a), w_1 \rangle$$

s.t. $w_h = \theta_h +$ LSVI-est. $\hat{w}_h$, $\|\theta_h\|_{\Lambda_h}^2 \leq \alpha_h^2$ & $\|w_h\|_2^2 \leq (\rho_h^w)^2$

$\theta_{t+1} = \theta_t + \eta w_t$

$\pi_h(\theta_{t+1})(a|s) \propto \pi_h(\theta_t)(a|s)\, e^{\eta \langle \phi(s,a), w \rangle \hat{Q}_h(s,a)}$

requiring Π_{soft} Bellman restricted closedness

Golowich & Moitra (2024) (not peer-reviewed)

lin. Bellman completeness : low inherent Bellman err.

$B_h = \{ w \in \mathbb{R}^d : |\langle \phi(s_h, a_h), w \rangle| \leq 1, \forall (s_h, a) \}$

A1: $$\sup_{w_{h+1} \in B_{h+1}} \inf_{w_h \in B_h} \sup_{(s,a) \in S_h \times A} \left| \langle \phi(s,a), w_h \rangle - \mathbb{E}_{s_{h+1}} [r_h(s,a) + \max_{a'} \langle \phi(s_{h+1}, a'), w_{h+1} \rangle] \right| \leq \frac{\varepsilon_{BE}}{2}$$

Thm: $\forall \pi \in \Pi^{\sigma}$ (pertubed linear), $h \in [H-1]$,

$$\sup_{w_{h+1} \in \mathbb{R}^d} \inf_{w_h} \sup_{(s,a)} \left| \langle \phi(s_h, a), w_h \rangle - \mathbb{E}_{\substack{s_{h+1} \\ a_{h+1} \sim \pi}} [\langle \phi(s_{h+1}, a_{h+1}), w_{h+1} \rangle] \right| \leq \tilde{O}\left(\|w\| d^{3/2} (\sqrt{d} + \tfrac{1}{\sigma}) \varepsilon_{BE} \right)$$

Def: $\sigma > 0$, $h \in [H]$, $\pi_{h,w,\sigma}(s)(a) = \mathbb{P}_{\theta \sim \mathcal{N}(w, \sigma^2 I_d)} (a \in \operatorname{argmax} \langle \theta, \phi(s,a) \rangle)$

Thm (informal): A2: $r_h(s,a) = \langle \phi(s,a), \theta_h^r \rangle$ determ. & lin. reward

$\|\phi\| \leq 1$, $\|\theta_h^r\| \leq 1$, $\|w\| \leq B$

$\varepsilon_{BE} \leq \iota_0 (BH)^{-2} d^{-3}$ A3: $C(\pi^*, \mathcal{D})$

$\exists$ algo. outputs $\hat{\pi}_T$, w.p. $\geq 1-\delta$, $V_1^*(s_1) - V_1^{\hat{\pi}_T}(s_1) \leq \tilde{O}\left(d^{3/2} BH \sqrt{\varepsilon_{BE}} + \frac{BHd}{\sqrt{n}} \right) (H + C(\pi^*, \mathcal{D}))$

$$+ \sqrt{\frac{64 B^2 H^2 \sqrt{d}}{T}}$$