Marginalized Importance Sampling (MIS)

$\rho_{1:H} \sim (\pi/\pi_b)^H$

"curse of horizon" [Liu et al '18]

- IS: exponential variance unless $\pi \approx \pi_b$.
- FQE. (policy-eval ver. of FQI).

$$f_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{|D|} \sum_{(s,a,r,s')} \left(f(s,a) - r - \gamma f_k(s', \pi) \right)^2$$

- $\forall f \in \mathcal{F}, \; \mathcal{T}^\pi f \in \mathcal{F}$.

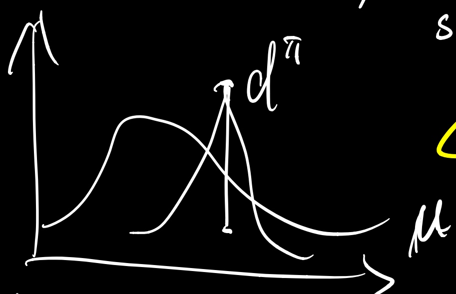
$\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{F}$. for FQI.

$\Rightarrow$ w/ $\mathcal{F}$ bounded complexity. poly sample size.

$$\| f_{k+1} - \mathcal{T}^\pi f_k \|_{2,\mu} \le \varepsilon$$

$(s,a) \sim \mu$, $r \sim R(s,a)$
$s' \sim P(\cdot|s,a)$.

$$\Rightarrow \left\| \frac{d^\pi}{\mu} \right\|_\infty \le C.$$



$$\Rightarrow \left| J(\pi) - \mathbb{E}_{s \sim d_0}[f_k(s, \pi)] \right|$$

$$\le \text{poly}(C) \cdot \varepsilon.$$

"realizability."

- Problem w/ FQE: $\boxed{Q^\pi \in \mathcal{F}}$ is insufficient
- Is there method w/ monotone func-approx error?
- Is there anything we can do. if we only have $Q^\pi \in \mathcal{F}$.

$\Rightarrow$ MIS address both questions.

Dai, Nachum... "DICE", "M_L".

"density estimation"....

$MDP = (S, A, P, R, \gamma, s_0)$ d_0.

W.t. learn $J(\pi) = Q^\pi(s_0, \pi) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi}[r]$. $r \sim R(s,a)$

have data: $(s,a) \sim \mu$, $r \sim R(s,a)$, $s' \sim P(\cdot|s,a)$

"Eval error Lemma for q":

TD/ Bellman error.

$\forall q.$ $J(\pi) - q(s_0, \pi) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi}[r + \gamma q(s', \pi) - q(s,a)]$.

Prof: $d^\pi = (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} d_t^\pi \longrightarrow$ dist. of (s_t, a_t) under policy π.

$$\begin{aligned} RHS = &\ \mathbb{E}_{(s,a)\sim d_1^\pi}[r + \cancel{\gamma q(s', \pi)} - \boxed{q(s,a)}] \\ &+ \gamma\, \mathbb{E}_{(s,a)\sim d_2^\pi}[r + \cancel{\gamma q(s', \pi)} - \cancel{q(s,a)}] \\ &+ \gamma^2 \ \cdots \end{aligned}$$

$(s,a) \sim d_1^\pi$
$s' \sim P(\cdot|s,a)$
$a' \sim \pi(\cdot|s')$

$\mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | \pi]$.

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}_{(s,a)\sim d_t^\pi}[r] - \mathbb{E}_{(s,a)\sim d_1^\pi}[q(s,a)].$$

$$= J(\pi) - q(s_0, \pi) = LHS. \quad \boxed{\checkmark}$$

Alt. proof: $J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi}[r]$.

the remaing: $0 = q(s_0, \pi) + \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi}[\gamma q(s', \pi) - q(s,a)]$

"Bellman eq for occupancy"

$$d^\pi = \begin{array}{|c|} \hline d_1^\pi \\ \hline \gamma d_2^\pi \\ \hline \gamma^2 d_3^\pi \\ \hline \gamma^3 d_4^\pi \\ \hline \end{array} \longleftrightarrow \begin{array}{|c|} \hline d_1^\pi \\ \hline \gamma d_2^\pi \\ \hline \gamma^2 d_3^\pi \\ \hline \vdots \\ \hline \end{array}$$

Goal: find q s.t. minimize

$$\min_q \left| J(\pi) - q(s_0,\pi) \right| = \left| \frac{1}{1-\gamma} \mathbb{E}_{d^\pi} [r + \gamma q(s',\pi) - q(s,a)] \right|$$

$$= \left| \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\mu} \left[\boxed{\frac{d^\pi(s,a)}{\mu(s,a)}} (r + \gamma q(s',\pi) - q(s,a)) \right] \right|$$

$w^{\pi/\mu}(s,a)$ "marginalized imp. weight" "density ratio"

Assume: $w^{\pi/\mu} \in \mathcal{W}$

$$\le \sup_{w\in\mathcal{W}} \frac{1}{1-\gamma} \left| \mathbb{E}_\mu [w(s,a)(r + \gamma q - q)] \right|$$

side. comment.

$\mathbb{E}_q \left[\frac{p(x)}{q(x)} f(x) \right]$, suppose. $y = g(x)$, $f(x) = f'(y)$.

$\mathbb{E}_q \left[\frac{p_Y(y)}{q_Y(y)} f'(y) \right]$

in general:

$$\mathrm{Var}\left[\frac{p_Y(y)}{q_Y(y)} \right] \le \mathrm{Var}\left[\frac{p(x)}{q(x)} \right]$$

$$\rho_{1:H} = \prod \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} = \underbrace{\frac{P^\pi(s_1,a_1,s_2,a_2,\cdots,s_H,a_H)}{P^{\pi_b}(s_1,a_1,s_2,a_2,\cdots,s_H,a_H)}}$$

$$w^{\pi/\mu} = \frac{P^\pi(s_t,a_t)}{P^{\pi_b}(s_t,a_t)} \leftarrow \text{or arbitrary } \mu.$$

in our ctx.

$X = s_1, a_1, s_2, a_2, \dots, s_H, a_H.$

$g(x) = s_t, a_t.$

$$|J(\pi) - q(s_0,\pi)| \le \sup_{w\in\mathcal{W}} \frac{1}{1-\gamma} \left| \mathbb{E}_\mu [w(s,a)(r + \gamma q(s',\pi) - q(s,a))] \right|$$

Alg. over $\mathcal{Q}$ class assuming $w^{\pi/\mu} \in \mathcal{W}$ [can relax to $w^{\pi/\mu} \in \mathrm{conv}(\mathcal{W})$].

$$\arg\min_{q\in\mathcal{Q}} \sup_{w\in\mathcal{W}} \left| \mathbb{E}_\mu [w \cdot (r + \gamma q - q)] \right|$$

"MQL [Uehara et al '20]"

Why relaxation from $w^{\pi/\mu}$ to $\sup_{w \in \mathcal{W}}$ make sense?

→ Ideally $q = Q^\pi$ → "tight relaxation"

for this func. loss: $\left| \sup_{w \in \mathcal{W}} \frac{1}{1-\gamma} \mathbb{E}_\mu [w \cdot (r + \gamma q(s', \pi) - q(s,a))] \right|$ $= 0$.

$\mathbb{E}[\cdot \mid s,a] = 0$.

→ MQL: $\boxed{\begin{array}{l} \text{when } w^{\pi/\mu} \in \mathcal{W} \\ \text{when } Q^\pi \in \mathcal{Q} \end{array}}$ valid upper bound of $|J(\pi) - q(s_0, \pi)|$

upper bound can be minimized to 0.

→ "q generator" → "w discriminator".

"MWL" "w generator" "q discriminator".

---

find q : s.t.

Bellman error.

→ $\mathbb{E}_\mu [w(s,a) \cdot (q(s,a) - r - \gamma q(s', \pi))] = 0$.

different: $\| q - \mathcal{T}^\pi q \|^2_{2,\mu}$.

$= \mathbb{E}_\mu \left[\left(q(s,a) - \mathbb{E}_{r, s' \mid s,a} [r + \gamma q(s', \pi)] \right)^2 \right]$.

LSTD.
avg Bellman err.
compare to sq. ptwise.
alt. view of MQL.

Want to learn w. s.t. $\quad J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_\mu[w^{\pi/\mu} \cdot r]$.

$$\underline{J(\pi)} - \frac{1}{1-\gamma}\mathbb{E}[\underline{w} \cdot r]$$

"Eval error lemma on w"

$$= \underline{Q^\pi(s_0,\pi)} + \frac{1}{1-\gamma}\mathbb{E}[w(s,a) \cdot (\gamma Q^\pi(s',\pi) - Q^\pi(s,a))]. \quad \triangle$$

Proof: $\frac{1}{1-\gamma}\mathbb{E}_\mu\left[w(s,a) \cdot \left(\underbrace{r + \gamma Q^\pi(s',\pi) - Q^\pi(s,a)}_{\mathbb{E}[\cdot \mid s,a] = 0.}\right)\right] = 0$.

---

find w. s.t. $\left|J(\pi) - \frac{1}{1-\gamma}\mathbb{E}_\mu[w(s,a) \cdot r]\right| \approx 0$.

$\boxed{\text{Assume } Q^\pi \in \mathcal{Q}.}$

$$\le \sup_{q \in \mathcal{Q}} \left|\underbrace{\left(q(s_0,\pi) + \frac{1}{1-\gamma}\mathbb{E}_\mu[w \cdot (\gamma q - q)]\right)}_{\triangle}\right|$$

"MWL": $\arg\min_{w \in \mathcal{W}} \sup_{q \in \mathcal{Q}}$ $\quad J(\pi) \in \frac{1}{1-\gamma}\mathbb{E}_\mu[\hat{w} \cdot r] \pm \sup_q |L_w(\hat{w}, q)|$

Relaxation is "tight": $w = \underline{w^{\pi/\mu}}$.

$$\forall q. \quad \underline{q}(s_0,\pi) + \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[(\gamma q(s',\pi) - q(s,a))].$$

$$= \langle q, \underbrace{d_0 \times \pi - \frac{1}{1-\gamma}d^\pi + \frac{\gamma}{1-\gamma}((P^T d^\pi) \times \pi)}_{= \vec{0}. \text{ b/c Bellman eq for } d^\pi} \rangle$$

$= 0.$

So,

if $Q^\pi \in \mathcal{Q} \Rightarrow$ valid upper bound.

if $w^{\pi/\mu} \in \mathcal{W} \Rightarrow$ minimize upper bound. (to 0).

$$J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_\mu[w \cdot r] + Q^\pi(s_0,\pi) + \frac{1}{1-\gamma}\mathbb{E}_\mu[w(\gamma Q^\pi(s',\pi) - Q^\pi(s,a))].$$

$$J(\pi) = q(s_0,\pi) + \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[r + \gamma q(s',\pi) - q(s,a)].$$

Define $L(w,q) = \frac{1}{1-\gamma}\mathbb{E}_\mu[w \cdot r] + q(s_0,\pi) + \frac{1}{1-\gamma}\mathbb{E}_\mu[w(\gamma q(s',\pi) - q(s,a))].$

$J(\pi) = L(w, Q^\pi) = L(w^{\pi/\mu}, q)$ $\forall w, q.$

$\boxed{Q^\pi \in \mathcal{Q}}$ then:

$\forall w.$ $J(\pi) = L(w, Q^\pi) \le \sup_{q\in\mathcal{Q}} L(w,q)$

$\ge \inf_{q\in\mathcal{Q}} L(w,q)$

$J(\pi) \in [\inf_q L(w,q), \sup_q L(w,q)]$ $\forall w.$

$$\Rightarrow \sup_{w\in\mathcal{W}} \inf_{q\in\mathcal{Q}} L(w,q) \le \boxed{J(\pi)} \le \inf_{w\in\mathcal{W}} \sup_{q\in\mathcal{Q}} L(w,q) \quad \checkmark$$

$\boxed{w^{\pi/\mu} \in \mathcal{W}}$

$\forall q.$ $\inf_{w\in\mathcal{W}} L(w,q) \le J(\pi) = L(w^{\pi/\mu}, q) \le \sup_{w\in\mathcal{W}} L(w,q)$

$$\Rightarrow \sup_{q\in\mathcal{Q}} \inf_{w\in\mathcal{W}} L(w,q) \le J(\pi) \le \inf_{q\in\mathcal{Q}} \sup_{w\in\mathcal{W}} L(w,q). \quad \checkmark$$

Sion's minimax thm: b/c $L(w,q)$ is convex-concave in w/q. and $\mathcal{W}$ & $\mathcal{Q}$ are convex.

$$\Rightarrow \sup_q \inf_w L(w,q) = \inf_w \sup_q L(w,q).$$

Why misspecification? $w^{\pi/\mu} \notin \mathcal{W}$ $Q^\pi \notin \mathcal{Q}$.

d^π

μ $\mu = d^{\pi_b}$.

How about learning? For OPE. we try to learn Q^π

For learning, we try to learn Q^*. $\max_{a'} Q^*(s',a')$

$\Rightarrow \forall w.$ $\mathbb{E}_\mu \left[w(s,a) \cdot \left(Q^*(s,a) - r - \gamma V_{Q^*}(s') \right) \right] = 0.$

"MABO"

$$\underset{g \in \mathcal{Q}}{\operatorname{argmin}} \ \boxed{\max_{w \in \mathcal{W}}} \left| \mathbb{E}_\mu \left[w \cdot \left(g(s,a) - r - \gamma V_g(s') \right) \right] \right|.$$

$\hookrightarrow$ output π_g. $\leftarrow$ when near-optimal?

Func-approx: ① $Q^* \in \mathcal{Q}$.

② $\forall \pi_g$ s.t. $g \in \mathcal{Q}$, $\frac{d^{\pi_g}}{\mu} \in \mathcal{W}$.

Lemma: $\forall g$.

$$J(\pi^*) - J(\pi_g) \le \frac{1}{1-\gamma} \left(\mathbb{E}_{d^{\pi^*}}[\underbrace{\mathcal{T}g - g}] + \mathbb{E}_{d^{\pi_g}}[g - \mathcal{T}g] \right).$$

$r + \gamma g(s',\pi) - g(s,a)$

Proof: Recall eval error lemma:

$$\forall \pi, \quad J(\pi) - g(s_0, \pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^\pi}[r + \gamma g(s',\pi) - g(s,a)].$$

$$J(\pi^*) - J(\pi_g) = \underbrace{J(\pi^*) - g(s_0, \pi_g)} + \underbrace{g(s_0, \pi_g) - J(\pi_g)}.$$

$$\le J(\pi^*) - g(s_0, \pi^*) + g(s_0, \pi_g) - J(\pi_g).$$

$\downarrow$ eval π^* use g. $\downarrow$ eval π_g use g.

$$= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi^*}}[r + \gamma g(s', \pi^*) - g(s,a)] - \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_g}}[r + \gamma g(s', \pi_g) - g(s,a)]$$

V_q.

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi^*}}[r+\gamma q(s',\pi_q)-q(s,a)]-\frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_q}}[\mathcal{T}q-q].$$

$$=\frac{1}{1-\gamma}\mathbb{E}_{d^{\pi^*}}[\mathcal{T}q-q]-\frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_q}}[\mathcal{T}q-q]. \quad \checkmark \quad \blacksquare$$

$$\arg\min_{w}\max_{q}\,|\cdots|$$
