OPE

- IS: exponential variance. ← can we avoid?
  "curse of horizon"

- FQE (policy-eval ver. of FQI).

FQI: $\max_a f_k(s,a)$ ⇓

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

Require: $\forall f\in\mathcal{F},\ \mathcal{T}^\pi f \in \mathcal{F}$

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

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

⇒ guarantee poly sample comp. for FQE.

- Can we make only realizability assumptions?

Marginalized IS

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

Then: if we know:

$$w^\pi(s,a) := \frac{d^\pi(s,a)}{\mu(s,a)} \quad (\text{assume} \le C)$$

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

in IS: $\dfrac{P^\pi(s_1,a_1,r_1,\dots,s_h,a_h,r_h)}{P^{\pi_b}(s_1,a_1,r_1,\dots,s_h,a_h,r_h)}\cdot r_h.$

in MIS: $\dfrac{P^\pi(s_h,a_h,r_h)}{P^{\pi_b}(s_h,a_h,r_h)}\cdot r_h.$

$w^\pi(s,a)$ is unknown — must "learn" using func. approx.

MQL] Try to learn $q \approx Q^\pi$ w/ the help of w^π

$$\forall q.\quad J(\pi) - \mathbb{E}_{s\sim d_0}[q(s,\pi)]$$

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

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

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

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

$w^\pi \in \mathcal{W}$ ← "realizability."

$$\leq \max_{w\in\mathcal{W}} \frac{1}{1-\gamma}\left|\mathbb{E}_\mu\left[w(s,a)\cdot\left(r+\gamma q(s',\pi) - q(s,a)\right)\right]\right|.$$

$L_q(w,q)$.

$$\hat{q} = \boxed{\operatorname*{argmin}_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} L_q(w,q).} \Rightarrow \hat{J}(\pi) = \mathbb{E}_{d_0}[\hat{q}].$$

Consistent estimator of $J(\pi)$ if $w^\pi \in \mathcal{W}$, $Q^\pi \in \mathcal{Q}$.

- When $w^\pi \in \mathcal{W}$, $\left|J(\pi) - \mathbb{E}_{d_0}[g]\right| \le \max_{w \in \mathcal{W}} L_g(w, g)$
- For $g = Q^\pi$, $\max_{w \in \mathcal{W}} L_g(w, g) = 0$.

(if $Q^\pi \in \mathcal{Q}$, $\min_{g \in \mathcal{Q}} \max_{w \in \mathcal{W}} L_g(w, g) = 0$

$\Rightarrow$ argmin g is accurate in its OPE prediction.

Remarks: $w^\pi \in \mathcal{W}$ can be relaxed to $w^\pi \in \mathrm{conv}(\mathcal{W})$.

MWL Goal: learn $w \approx w^\pi$

i.e. learn w to minimize

$$\left|J(\pi) - \frac{1}{1-\gamma}\mathbb{E}_\mu[w \cdot r]\right| \quad (*).$$

$\mathbb{E}[\cdot \mid s,a] = R(s,a)$

$$= \left|\mathbb{E}_{s \sim d_0}[Q^\pi(s,\pi)] + \frac{1}{1-\gamma}\mathbb{E}_\mu\left[w(s,a)\left(\gamma Q^\pi(s',\pi) - Q^\pi(s,a)\right)\right]\right|$$

$Q^\pi \in \mathcal{Q}$

$$\le \max_{g \in \mathcal{Q}} \left|\mathbb{E}_{s \sim d_0}[g(s,\pi)] + \frac{1}{1-\gamma}\mathbb{E}_\mu\left[w(s,a)\left(\gamma g(s',\pi) - g(s,a)\right)\right]\right|$$

$= \max_{q \in \mathcal{Q}} L_w(w, q)$.

Bellman flow error of $\mu \cdot w$.

Verify this is tight upper bound:

$\forall q$, $L_w(w^\pi, q)$

$= \left| \mathbb{E}_{s \sim d_0}[q(s,\pi)] + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d^\pi}[\gamma q(s',\pi) - q(s,a)] \right|$

$= 0$. (Bellman flow eq for d^π).

$-q(s,a)$ w/ $(s,a) \sim \sum_{t=1}^{\infty} \gamma^{t-1} d_t^\pi$

$+q(s,a)$ w/ $(s,a) \sim d_1^\pi$.

$+q(s,a)$ w/ $(s,a) \sim \gamma P(d^\pi) \times \pi$

$= \sum_{t=2}^{\infty} \gamma^{t-1} d_t^\pi$.

MIS for learning good policy?

$\mathbb{E}_\mu[w \cdot (q - \mathcal{T}q)]$

Learn $q \approx Q^*$ s.t. π_q is a good policy.

$$(*) \quad \operatorname{argmin}_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} \frac{1}{1-\gamma} \left| \mathbb{E}_\mu \left[w(s,a) \cdot \left(q(s,a) - r - \gamma \max_{a'} q(s',a') \right) \right] \right|$$

When does this work? $Q^* \in \mathcal{Q}$.

$$J(\pi^*) - J(\pi_q) \le \frac{1}{1-\gamma}\Big(\mathbb{E}_{d^{\pi^*}}[\mathcal{T}q - q] + \mathbb{E}_{d^{\pi_q}}[q - \mathcal{T}q]\Big).$$

can't guarantee $q = \mathcal{T}q$ on all states.

Implies $\Rightarrow$ $(*)$ learns good policy if $w^{\pi_q} \in \mathcal{W}, \forall q \in \mathcal{Q}$.

Proof: Recall $\forall \pi, q$, $J(\pi) - \mathbb{E}_{d_0}[q(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[\mathcal{T}^\pi q - q]$.

$$J(\pi^*) - J(\pi_q) \le J(\pi^*) - \mathbb{E}_{s\sim d_0}[q(s,\pi^*)] + \mathbb{E}_{d_0}[q(s,\pi_q)] - J(\pi_q)$$

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

$q \approx Q^*$ $\begin{cases} q \approx Q^{\pi^*} \\ q \approx Q^{\pi_q} \end{cases}$

$$\le \frac{1}{1-\gamma}\Big(\mathbb{E}_{d^{\pi^*}}[\mathcal{T}q - q] + \mathbb{E}_{d^{\pi_q}}[q - \mathcal{T}q]\Big).$$

Note: if $Q^* \in \mathcal{Q}$ and $\forall q \in \mathcal{Q}$, $w^{\pi_q} \in \mathcal{W}$.

$$\Big|\mathbb{E}_{d^{\pi^*}}[\mathcal{T}q - q] + \mathbb{E}_{d^{\pi_q}}[q - \mathcal{T}q]\Big| \le 2 \cdot \max_{w \in \mathcal{W}} \Big|\mathbb{E}_\mu[w \cdot (q - \mathcal{T}q)]\Big|.$$