$D = \{s, a, s'\}$ $\quad s,a \sim \rho, \quad s' \sim P(\cdot \mid s,a)$

Fitted Q-Iteration:

Given $\mathcal{F} \subseteq S \times A \mapsto [0, \frac{1}{1-\gamma}]$

Assumptions:

① Bell-completion: $\forall f \in \mathcal{F}: \quad \mathcal{T} f \in \mathcal{F}$ ($\mathcal{T}$: Bell-operator)

② $\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} \leq C < \infty$

Min-Max Approach For PE

① Coverage condition: $\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} \leq C < \infty$

Function classes: $\mathcal{F} \subseteq S \times A \mapsto [0, \frac{1}{1-\gamma}], \quad (Q^{\pi} \in \mathcal{F})$

$\mathcal{W} \subseteq S \times A \mapsto [-C, C] \quad (\frac{d^{\pi}}{\rho} \in \mathcal{W})$

② Realizability: $Q^{\pi} \in \mathcal{F}, \quad \frac{d^{\pi}}{\rho} \in \mathcal{W}$

③ symmetric: $w \in \mathcal{W}, \Rightarrow -w \in \mathcal{W}$

Algorithm:

Define: $\hat{l}(w, Q) := \frac{1}{N} \sum_{i=1}^{N} w(s_i, a_i) \left(Q(s_i, a_i) - r(s_i, a_i) - \gamma Q(s_i', \pi(s_i')) \right)$

$l(w, Q) = \mathop{E}_{s,a \sim \rho} \left[w(s,a) \left(Q(s,a) - r(s,a) - \gamma \mathop{E}_{s' \sim P(s,a)} Q(s', \pi(s')) \right) \right]$

Min-max PE:

$\hat{Q} = \mathop{\text{argmin}}_{Q \in \mathcal{F}} \max_{w \in \mathcal{W}} \hat{l}(w, Q)$

$\hat{V}^{\pi} := \hat{Q}(s_0, \pi(s_0)) \quad (s_0 \text{ is fixed})$

Goal: $|\hat{V}^{\pi} - V^{\pi}| \le \frac{1}{\sqrt{n}} \cdot \ln(|W||F|/\delta) \cdot \text{poly}\left(\frac{1}{1-\gamma}, c\right)$

Analysis:

① Uniform convergence: (Hoeffding + Union Bound)

$\forall w \in W,\ Q \in F$:

$$|\hat{\ell}(w, Q) - \ell(w, Q)| \le \underbrace{\sqrt{\frac{\ln(|W||F|/\delta)}{N}} \cdot \text{poly}\left(\frac{1}{1-\gamma}, c\right)}_{\delta}$$

whp

② min-max

$$\min_{Q} \underbrace{\left[\max_{w} \hat{\ell}(Q, w)\right]}_{\text{loss for } Q}$$

$\forall Q$:

$$\left|\max_{w} \hat{\ell}(Q, w) - \max_{w} \ell(Q, w)\right| \le \max_{w} \left|\underbrace{\hat{\ell}(Q, w) - \ell(Q, w)}\right| \le \delta$$

Recall: $\hat{Q} = \arg\min_{Q} \max_{w} \hat{\ell}(Q, w)$ (*)

$\max_{w} \ell(w, Q^{\pi}) = 0$ (*,*)

$$\max_{w} \ell(w, \hat{Q}) \le \max_{w} \hat{\ell}(w, \hat{Q}) + \delta$$

$$\le \max_{w} \hat{\ell}(w, Q^{\pi}) + \delta \quad (*)$$

$$\le \max \ell(w, Q^{\pi}) + 2\delta$$

$$= 2\delta \qquad (*,*)$$

(3) $\max_w \ell(w, \hat{Q}) \le 2\delta \Rightarrow |\hat{V}^\pi - V^\pi|$

$$\max_{w \in W} \mathop{E}_{\text{samp}} \left(w(s,a)\left(\hat{Q}(s,a) - r(s,a) - \gamma \mathop{E}_{s' \sim P_{sa}} \hat{Q}(s', \pi(s'))\right)\right) \le 2\delta \quad (a)$$

By Symmetric condition on W:

$$\max_{w \in W} \mathop{E}_{\text{samp}} \left[w(s,a)\left(-\hat{Q}(s,a) + r(s,a) + \gamma \mathop{E}_{s' \sim P_{sa}} \hat{Q}(s', \pi(s'))\right)\right] \le 2\delta \quad (b)$$

By $w^\pi := \frac{d^\pi}{\rho} \in W$, & (a)+(b)

$$\left| \mathop{E}_{\text{samp}} \left[w^\pi(s,a)\left(\hat{Q}(s,a) - r(s,a) - \gamma \mathop{E}_{s' \sim P_{sa}} \hat{Q}(s', \pi(s'))\right)\right]\right| \le 2\delta$$

$$\min_{\hat{Q}} \sum_{i=1}^{n} \left(\underbrace{\hat{Q}(s,a)}_{A} - r(s,a) - \gamma \underbrace{\hat{Q}(s', \pi(s'))}_{A}\right)^2 / n \Leftarrow \checkmark$$

is not unbiased wrt:

$$\mathop{E}_{\text{samp}} \left[\left(\hat{Q}(s,a) - r(s,a) - \gamma \mathop{E}_{s' \sim P(sa)} \hat{Q}(s', \pi(s'))\right)^2 \right] \checkmark$$

$$\left| \mathop{E}_{s,a \sim d^\pi} \left(\hat{Q}(s,a) - r(s,a) - \gamma \mathop{E}_{s' \sim P_{sa}} \hat{Q}(s', \pi(s'))\right)\right| \le 2\delta \quad (**)$$

$$d^\pi(s,a) = (1-\gamma)\,\delta(s_0)\,1(a = \pi(s_0)) + \gamma \mathop{E}_{\tilde{s},\tilde{a} \sim d^\pi} P(s \mid \tilde{s}, \tilde{a})\, 1(a = \pi(s)) \quad (*)$$

$$E_\pi \hat{Q}(s,a) - \gamma E \quad E \quad \hat{Q}(s', \pi(s'))$$

s and a $\sim$ $s, a \sim d^{\pi}$ simply

$$= (1-\gamma)\,\widehat{Q}(s_0, \pi(s_0)) \quad (*)$$

$$= (1-\gamma)\,\hat{V}^{\pi}$$

Together w/ $(**)$

$$\Rightarrow \left| (1-\gamma)\hat{V}^{\pi} - (1-\gamma)V^{\pi} \right| \le \epsilon$$

$$:= \mathop{E}_{s,a \sim d^{\pi}} r(s,a)$$

□

Summary:

$$\left| \hat{V}^{\pi} - V^{\pi} \right| \le \frac{1}{\sqrt{n}} \cdot \sqrt{\ln(|M||F|/\delta)} \cdot \mathrm{poly}\left(\frac{1}{1-\gamma}, c\right)$$