Batch RL + Fitted Q-iteration

a) Coverage Assumptions
b) Function Approx Assumptions

MDP: $S, A, P, R: S \times A \to [0,1]$, $\gamma \in [0,1)$, $d_0 \in \Delta(S)$

$V_{max} := \frac{1}{1-\gamma}$

Data: $\{(x_i, a_i, r_i, x_i')\}_{i=1}^n$ $\quad (x_i, a_i) \overset{iid}{\sim} \mu \in \Delta(S \times A)$, $r_i \sim R(x_i, a_i)$, $x_i' \sim P(\cdot | x_i, a_i)$

Goal: Output $\hat{\pi}$ s.t. with prob $1-\delta$: $V(\hat{\pi}) \geq V^* - \varepsilon$

Example: $\gamma = 0$, $d_0 = \delta_{x_0}$ $\Rightarrow$ MAB problem

$\to \mu = d_0 \circ \delta_a$, cannot learn reward for $a' \neq a$

$\Rightarrow$ Cannot find π^*

$\Rightarrow$ Must make coverage assumption

Assmp 1: (Unif concentrability) $\exists\, C_{unif} < \infty$ s.t. $\forall x, a: \frac{1}{\mu(x,a)} \leq C_{unif}$

$\to \approx \frac{1}{|S| \times |A|}$

$\Rightarrow |S| \times |A| \leq C_{unif}$ $\Rightarrow$ Tabular Setting

Prop 1: Under Assmp 1, $\exists$ an alg s.t.

$$V^* - V(\hat{\pi}) \leq \tilde{O}\left(\frac{1}{(1-\gamma)^2} \left(\sqrt{\frac{C_{unif}}{n}} + V_{max} \gamma \left(|S| \sqrt{\frac{C_{unif}}{n}} \right) \right) \right)$$

($\sqrt{\frac{C_{unif}}{n}} = \Delta_R$, $|S|\sqrt{\frac{C_{unif}}{n}} = \Delta_P$)

$\tilde{O}$ hiding $\log(|S||A|/\delta)$ factors

Sketch: n samples, $\frac{n}{C_{unif}}$ samples from each (x,a)

$$\hat{R}(x,a) = \frac{\sum_i \mathbb{1}\{(x_i,a_i)=(x,a)\} r_i}{\sum_i \mathbb{1}\{(x_i,a_i)=(x,a)\}} \qquad \hat{P}(x'|x,a) = \frac{\sum_i \mathbb{1}\{(x_i,a_i,x_i')=(x,a,x')\}}{\sum_i \mathbb{1}\{(x_i,a_i)=(x,a)\}}$$

Concentration

$$\forall x,a: \quad |\hat{R}(x,a) - R(x,a)| \lesssim \sqrt{\frac{C_{unif}}{n}} =: \Delta_R \leftarrow \text{Hoeffding + Union}$$

$$\forall x,a: \quad \|\hat{P}(\cdot|x,a) - P(\cdot|x,a)\|_{TV} \lesssim |S| \sqrt{\frac{C_{unif}}{n}} =: \Delta_P$$

$\hat{\pi}$, opt in $(\hat{P},\hat{R})$

$$V^* - V(\hat{\pi}) = \sum_{h\ge 0} \gamma^h \mathbb{E}_{d_{h,\hat{\pi}}}\left[Q^*(x,\pi^*(x)) - Q^*(x,\hat{\pi}(x)) \right] \leftarrow \text{PDL}$$

$$\pm \hat{Q}(x,\pi^*(x)) \leftarrow$$

$$\pm \hat{Q}(x,\hat{\pi}(x)) \leftarrow$$

$$\hat{Q}(x,\hat{\pi}(x)) \ge \hat{Q}(x,\pi^*(x))$$

$$\le \sum_h \gamma^h \mathbb{E}_{d_{h,\hat{\pi}}}\left[|Q^*(x,\pi^*(x)) - \hat{Q}(x,\pi^*(x))| + |\hat{Q}(x,\hat{\pi}(x)) - Q^*(x,\hat{\pi}(x))| \right] \leftarrow$$

$$\mathbb{E}_{\nu}\left[|Q^*(x,a) - \hat{Q}(x,a)|\right] \le \mathbb{E}_{\nu}\left[|R(x,a) - \hat{R}(x,a)|\right] \lesssim \Delta_R$$

$$+ \mathbb{E}_{\nu}\left[\gamma \left| \mathbb{E}_{x'\sim P_{x,a}} V^*(x') - \mathbb{E}_{x'\sim \hat{P}_{x,a}} \hat{V}(x') \right|\right]$$

$= R + P V^*(x')$ → $\hat{R} + \hat{P}\hat{V}$

$$\gamma \mathbb{E}_{\nu}\left[\left| \mathbb{E}_{x'\sim P_{x,a}} V^*(x') - \hat{V}(x') \right| \right] + \gamma \mathbb{E}_{\nu}\left[\left| \mathbb{E}_{P} \hat{V}(x') - \mathbb{E}_{\hat{P}} \hat{V}(x') \right| \right]$$

↳ $\gamma V_{\max} \Delta_P$

$$\gamma \mathbb{E}_{v}\left[\left|\mathbb{E}_{x' \sim P_{x,a}} V^*(x') - \hat{V}(x')\right|\right] + \gamma \mathbb{E}_{v}\left[\left|\mathbb{E}_{P} \hat{V}(x') - \mathbb{E}_{\hat{P}} \hat{V}(x')\right|\right]$$

$\hookrightarrow \gamma V_{max} \Delta_P$

$V^*(x') = \max_a Q^*(x,a)$

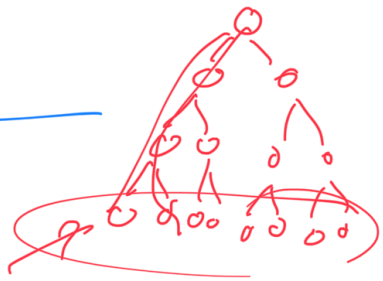
$\hat{V}(x') = \max_a \hat{Q}(x',a)$

$$|V^*(x') - \hat{V}(x')| \le |Q^*(x', \pi(x')) - \hat{Q}(x', \pi(x'))|$$

$$\text{error} \le \Delta_R + V_{max} \gamma \Delta_P + \gamma \cdot \text{error} \Rightarrow \text{error} \le \frac{1}{1-\gamma}(\Delta_R + V_{max} \gamma \Delta_P)$$

Assmp 2: (Concentrability) $\exists C < \infty$ s.t.,

$\forall$ nonstationary π, h, x,a $\quad \frac{d_h^{\pi}(x,a)}{\mu(x,a)} \le C$



$\tilde{\mathcal{F}} : S \times A \to [0, V_{max}]$

FQI: Initialize $f_0 \in \mathcal{F}$ arbitrarily

Iterate:

$$f_k \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(f(x_i,a_i) - \underbrace{r_i - \gamma \max_{a'} f_{k-1}(x_i', a')}_{(\mathcal{T} f_{k-1})(x_i,a_i)} \right)^2$$

$f_k \approx (\mathcal{T} f_{k-1})(x_i,a_i) \in \tilde{\mathcal{F}}$

output π_{f_K} for the last K

Bellman op: $\mathcal{T}f : (x,a) \mapsto \mathbb{E}\left[r + \gamma \max_{a'} f(x',a') \mid x,a\right]$

Assmp 3: (Completeness) $\mathcal{T}\tilde{\mathcal{F}} \subseteq \tilde{\mathcal{F}}$

$\hookrightarrow = \{\mathcal{T}f : f \in \tilde{\mathcal{F}}\}$

$\Rightarrow Q^* \in \tilde{\mathcal{F}}$ $\qquad Q^* = \lim_{k \to \infty} \mathcal{T}^k f_0$ for any f_0 $\qquad |\mathcal{F}| < \infty$

Thm: The k^{th} iterate of FQI satisfies w.p. $\ge 1-\delta$

$$V^* - V(\pi_{f_k}) \le O\left(\frac{V_{max}}{(1-\gamma)^2}\sqrt{\frac{C \log(|\mathcal{F}|/\delta)}{n}}\right) + \frac{2\gamma^k V_{max}}{(1-\gamma)}$$

Thm: The k^{th} iterate of FQI satisfies w.p. $\geq 1-\delta$

$$V^* - V(\pi_{f_k}) \leq O\left(\frac{V_{max}}{(1-\gamma)^2} \sqrt{\frac{C \log(|F|/\delta)}{n}} \right) + \frac{2\gamma^k V_{max}}{(1-\gamma)}$$

$\to 0$ as $n \to \infty$; opt. term

Proof: $\|f\|_\nu^2 = \mathop{\mathbb{E}}_{x,a \sim \nu} f(x,a)^2 \quad \nu \in \Delta(S \times A)$

Concentration: w.p. $1-\delta$

$$\forall k \quad \|f_k - \mathcal{T} f_{k-1}\|_\mu^2 \leq O\left(\frac{V_{max}^2 \log(|F|/\delta)}{n} \right) =: \Delta_{SQ}$$

$$V^* - V(\pi_k) \leq \sum_{h=0}^{\infty} \gamma^h \mathop{\mathbb{E}}_{d_{h,\pi_k}} Q^*(x,\pi^*(x)) - Q^*(x,\pi_k(x))$$

$$\leq \sum_{h=0}^{\infty} \gamma^h \|Q^* - f_k\|_{2, d_{h,\pi_k} \circ \pi^*} + \gamma^h \|Q^* - f_k\|_{2, d_{h,\pi_k} \circ \pi_k}$$

$$\|Q^* - f_k\|_{2,\nu} \leq \|Q^* - \mathcal{T} f_{k-1}\|_{2,\nu} + \|\mathcal{T} f_{k-1} - f_k\|_{2,\nu}$$

Assump: $\forall x,a, \ \frac{\nu(x,a)}{\mu(x,a)} \leq C$

$\sqrt{\mathbb{E}_\nu (\dots)^2}$

$\leq \sqrt{C \cdot \mathbb{E}_\mu (\dots)^2} = \sqrt{C}\, \|\mathcal{T} f_{k-1} - f_k\|_{2,\mu} \leq \sqrt{C \cdot \Delta_{SQ}}$

$\mathcal{T} f_{k-1} = R + \gamma f_{k-1}$; $Q^* = R + \gamma Q^*$

$$\leq \gamma \sqrt{\mathbb{E}_\nu \left[\left| \mathop{\mathbb{E}}_{x' \sim P} \max_{a'} Q^*(x',a') - \max_{a'} f_{k-1}(x',a') \right|^2 \right]}$$

$\tilde{\pi} = \text{argmax}_a \{ \max_{a'} Q^*(x,a'), \max_{a'} f_{k-1}(x,a') \}$

$$\leq \gamma \|Q^* - f_{k-1}\|_{\nu \circ \tilde{\pi}}$$

$$\to \|Q^* - f_k\|_{2,\nu} \leq \frac{1}{1-\gamma} \sqrt{C \cdot \Delta_{SQ}} + \gamma^k V_{max}$$

Concentrability:

Completeness

Tables in Linear MDP

Concentrability:

If f is linear in ϕ, then

$\|f\|_\nu^2 = f^\top \Sigma_\nu f$ $\quad \Sigma_\nu = \mathbb{E}_{x,a,x' \sim \nu} \phi \phi^\top$

$\hookrightarrow \le C \|f\|_\mu^2$

$f^\top \Sigma_\nu f = f^\top \Sigma_\mu^{1/2} \Sigma_\mu^{-1/2} \Sigma_\nu \Sigma_\mu^{-1/2} \Sigma_\mu^{1/2} f$

$\mathrm{tr}\left(\Sigma_\mu^{-1/2} \Sigma_\nu \Sigma_\mu^{-1/2} \cdot \Sigma_\mu^{1/2} f f^\top \Sigma_\mu^{1/2} \right)$

$\le \lambda_{\max}\left(\Sigma_\mu^{-1/2} \Sigma_\nu \Sigma_\mu^{-1/2} \right) \cdot f^\top \Sigma_\mu f$

"Linear" Concentrability

$\forall \pi, h, \quad \lambda_{\max}\left(\Sigma_{\pi,h} \cdot \Sigma_\mu^{-1} \right) \le C_{lin}$

Thm (Sok): linear completeness + linear concentrability

$V^* - V(\pi_k) \le \mathrm{poly}\left(C_{lin}, d, \frac{1}{n} \right) + O\left(\frac{\gamma^k V_{max}}{1-\gamma} \right)$

Completeness

Very strong, holds in Linear MDP in LQR

Non-monotonic

$G \supseteq F$

$\mathcal{T} F \subseteq F \Rightarrow TG \subseteq G$

$Q^* \in F$ is monotone

Realizability only: [Xie, Jiang]

$V^* - V(\hat{\pi}) \le n^{-1/4} \cdot \log(|F|/\delta)$

Very strong concentrability assumption

$\mu(a|s) \ge \frac{1}{C_A} \quad \frac{P(s'|s,a)}{\mu(s')} \le C_S$

$\hookrightarrow \Rightarrow$ our concentrability assumption

Shan, Ruosong, Wang, Den Foster

Realizability only not enough w/ "weak" concentrability

F linear, "linear" concentrability need ~~exp(H)~~ samples $\exp(H)$

minimax approach

$\underset{f \in F}{\mathrm{argmin}} \max_{w \in W} \left| \mathcal{L}(f, w) \right|$

$\mathcal{L} = \mathbb{E}_\mu \left[w(x,a) \left(f(x,a) - r - \gamma \max_{a'} f(x', a') \right) \right]$

$w(x,a) \hookrightarrow \frac{\delta''(x,a)}{\mu(x,a)}$

Horizon Dependence

$\frac{1}{(1-\gamma)^2}$ dependence, $\frac{1}{1-\gamma}$ possible

Bellman Residual Minimization

Minimax approach

$$\operatorname{argmin}_{f \in F} \max_{w \in W} \left| \underbrace{L(f, w)}_{\hookrightarrow = \mathbb{E}_{\mu}\left[w(x,a) \left(f(x,a) - r - \gamma \max_{a'} f(x', a') \right) \right]} \right|$$

$w(x,a)$ ↳ $\frac{d^{\pi}(x,a)}{\mu(x,a)}$

Horizon Dependence

$\frac{1}{(1-\gamma)^2}$ dependence, $\frac{1}{1-\gamma}$ possible

Bellman Residual Minimization

$$\operatorname{argmin}_{f \in F} \max_{g \in F} \mathbb{E}_{\mu}\left(f - r - \gamma \max_{a'} f(x', a') \right)^2 - \mathbb{E}_{\mu}\left(g - r - \gamma \max_{a'} f \right)^2$$

noise level for "fs problem"

$$V(\pi) - V(\pi_f) \le \frac{1}{1-\gamma} \left(\mathbb{E}_{d_{\pi}}\left[\mathcal{T} f - f \right] + \mathbb{E}_{d_f}\left[f - \mathcal{T} f \right] \right)$$

$f_k - \mathcal{T} f_{k-1}$

What if we have poor coverage?

- Obviously cannot match V^*

Model based ⇒ realizability only is ok

(but model-based realizability is V. strong!)

Model-free ⇒ realizability only seems insufficient

(but much weaker)