MDP (tabular analysis) $M = (S, A, P, R, \gamma)$.

Protocol: for each (s,a) pair.

n iid samples of $s' \sim P(\cdot|s,a)$.

data $D_{s,a}$ and $r \sim R(\cdot|s,a)$.

$\hat{\pi}$ $\|V_M^* - V_M^{\hat{\pi}}\|_\infty \le \varepsilon$.

Alg: $\hat{R}(s,a) = \frac{1}{n} \sum_{r_i \in D_{s,a}} r_i$.

const.

$$\hat{P}(s'|s,a) = \frac{1}{n} \sum_{s_i' \in D_{s,a}} \mathbb{I}[s_i' = s']$$

r.v.

$$\hat{P}(\cdot|s,a) = \frac{1}{n} e_{s_i'}$$

$s_i' \to$ [0 ⋮ 1 ⋮ 0]

$\hat{M} = (S, A, \hat{P}, \hat{R}, \gamma)$, $\hat{\pi} = \pi^*_{\hat{M}}$

① Concentration: $\hat{M} \approx M$.

(w.p. $\ge 1-\delta$) $\forall s,a$, $|\hat{R}(s,a) - R(s,a)| \le \varepsilon_R$.

$\| \hat{P}(\cdot|s,a) - P(\cdot|s,a) \|_1 \le \varepsilon_P$.

② "error propagation analysis".

$$\| V_M^* - V_M^{\pi_{\hat{M}}^*} \|_\infty \le f_n(\varepsilon_R, \varepsilon_P)$$

$\forall (s,a) \in S \times A$, $|\hat{R}(s,a) - R(s,a)| \le R_{max} \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$.

w.p. $\ge 1-\delta$.

$\forall (s,a,s')$, w.p. $\ge 1-\delta$. $|\hat{P}(s'|s,a) - P(s'|s,a)| \le \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$.

Given total failure budget δ.

- all reward estm: $\delta/2$ → evenly over $S \times A$.
- all transition: $\delta/2$. → evenly over $S \times A \times S$.

⇓

w.p. $\ge 1-\delta$. | $\forall (s,a)$.

$$|\hat{R}(s,a) - R(s,a)| \le R_{max}\sqrt{\frac{1}{2n}\ln\frac{4|S\times A|}{\delta}}$$

$$\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \le |S|\sqrt{\frac{1}{2n}\ln\frac{4|S\times A\times S|}{\delta}}$$

"Simulation Lemma" $\forall \pi: S \to A$.

$R_{max}/(1-\gamma)$.

$$\|V_{\hat{M}}^{\pi} - V_{M}^{\pi}\|_\infty \le \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma \varepsilon_P V_{max}}{1-\gamma}$$

Δ.

$$\|V_M^{\pi_M^*} - V_M^{\pi_{\hat{M}}^*}\|_\infty \le \|V_M^{\pi_M^*} - V_{\hat{M}}^{\pi_M^*} + V_{\hat{M}}^{\pi_{\hat{M}}^*} - V_M^{\pi_{\hat{M}}^*}\|_\infty$$

$$\le 2 \cdot \max_{\pi: S\to A} \|V_M^{\pi} - V_{\hat{M}}^{\pi}\|_\infty.$$

Proof of sim. Lemma: $\forall s$.

$V^{\pi}(s) = Q^{\pi}(s,\pi)$

$$|V_M^{\pi}(s) - V_{\hat{M}}^{\pi}(s)|.$$

$$= |R(s,\pi) + \gamma \langle P(\cdot|s,\pi), V_M^{\pi}(\cdot)\rangle$$

$$\left| \hat{R}(s,\pi) - \gamma \langle \hat{P}(\cdot|s,\pi), V_{\hat{M}}^{\pi}(\cdot) \rangle \right|$$

$$\leq \varepsilon_R + \gamma \left| \langle P(\cdot|s,\pi), V_M^{\pi} \rangle - \langle \hat{P}(\cdot|s,\pi), V_{\hat{M}}^{\pi} \rangle \right.$$

$$\left. - \langle P(\cdot|s,\pi), V_{\hat{M}}^{\pi} \rangle + \langle P(\cdot|s,\pi), V_{\hat{M}}^{\pi} \rangle \right|$$

$$\leq \varepsilon_R + \gamma \left| \langle \hat{P}(\cdot|s,\pi) - P(\cdot|s,\pi), V_{\hat{M}}^{\pi} \rangle \right| \leq O(\varepsilon_P).$$

$$+ \gamma \left| \langle P(\cdot|s,\pi), V_M^{\pi} - V_{\hat{M}}^{\pi} \rangle \right|$$

$$\leq \gamma \cdot \| V_M^{\pi} - V_{\hat{M}}^{\pi} \|_{\infty}$$

Hölder's ineq: for $\|\cdot\|$ and $\|\cdot\|_*$.

$$|\langle u, v \rangle| \leq \|u\| \cdot \|v\|_*.$$

Example: $\|\cdot\|_p$ & $\|\cdot\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$.

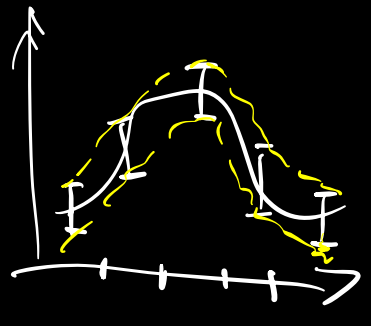
e.g. $p = q = 2$.

$p = 1, q = \infty$.

$$\left| \langle \hat{P}(\cdot|s,\pi) - P(\cdot|s,\pi),\ V_{\hat{M}}^{\pi} \rangle \right|$$

$$\leq \left\| \hat{P}(\cdot|s,\pi) - P(\cdot|s,\pi) \right\|_1 \cdot \left\| V_{\hat{M}}^{\pi} \right\|_\infty$$

$$\leq \varepsilon_p \cdot \frac{R_{max}}{1-\gamma} \to V_{max}.$$



p $(= P(\cdot|s,\pi))$ $\quad \hat{p}$ $(= \hat{P}(\cdot|s,\pi))$

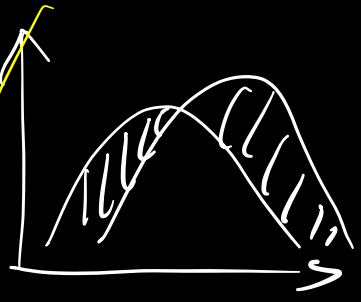
S. $\qquad S$.

$$\|p - \hat{p}\|_1 = \max_{u \in \{-1,1\}^S} u^T (p - \hat{p}).$$

$$u^T p - u^T \hat{p}$$

$$= u^T p - u^T \frac{1}{n} \sum_{s' \in D_{s,a}} e_{s'}$$

$e_{s'} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} s'$



$$= \mathbb{E}_{s' \sim p}[u(s')] - \frac{1}{n} \sum_{s' \in D_{s,a}} u(s').$$

$$\sum_s |p(s) - \hat{p}(s)|$$

$$= \sum_s \mathrm{sgn}[p(s) > \hat{p}(s)] (p(s) - \hat{p}(s))$$

$$= \max_{u \in \{-1,1\}^S} u^T (p - \hat{p})$$

w.p. $\geq 1-\delta$.

$$\leq 2 \cdot \sqrt{\frac{1}{n} \ln \frac{2}{\delta}}$$

w.p. $\geq 1-\delta$, $\|\hat{P} - P\|_1 \leq 2\sqrt{\frac{1}{2n} \ln \frac{2 \cdot 2^S}{\delta}} \simeq \sqrt{\frac{|S|}{n} \ln \frac{1}{\delta}}$

$2\left(\frac{\varepsilon_R^b}{1-\gamma} + \gamma \frac{\varepsilon_P^{\Phi} V_{max}}{1-\gamma} \right)$.

Alt. bound: $\forall f \in \mathbb{R}^{S\times A}$

$$\|V_M^* - V_M^{\pi_f}\|_\infty \le \frac{2\|f - Q_M^*\|}{1-\gamma}$$

$$\|V_M^* - V_M^{\pi_{\hat{M}}^*}\|_\infty \le \frac{2\|Q_{\hat{M}}^* - Q_M^*\|_\infty}{1-\gamma}$$

$$Q_M^*(s,a) - Q_{\hat{M}}^*(s,a)$$

$$= Q_M^*(s,a) - \hat{R}(s,a) - \gamma \langle \hat{P}(\cdot|s,a), V_{\hat{M}}^* \rangle$$

$$= Q_M^*(s,a) - \hat{R}(s,a) - \gamma \langle \hat{P}(\cdot|s,a), V_M^* \rangle$$

$$+ \gamma \langle \hat{P}(\cdot|s,a), V_M^* \rangle - \gamma \langle \hat{P}(\cdot|s,a), V_{\hat{M}}^* \rangle$$

$$Q_M^*(s,a) - \frac{1}{n}\sum_{r_i \in D_{s,a}} r_i - \frac{1}{n}\Big\langle \sum_{s_i' \in D_{s,a}} e_{s_i'}, V_M^* \Big\rangle$$

$$= Q_M^*(s,a) - \frac{1}{n}\sum_{(r_i, s_i') \in D_{s,a}} \underbrace{\left(r_i + \gamma V_M^*(s_i')\right)}_{[0,\, V_{max}]}.$$

w.p. $\ge 1-\delta$.

$\leq V_{max} \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}.$

$\leq 2 \max_{\pi} \| V_\gamma^\pi - V_{\gamma'}^\pi \|_\infty$

$Q_\gamma^* - Q_{\gamma'}^*$

$\leq Q_\gamma^{\pi_\gamma^*} - Q_{\gamma'}^{\pi_\gamma^*}$

Lemma: $\forall f \in \mathbb{R}^S$. fix. d_0.

$$\mathbb{E}_{s\sim d_0}[f(s_0)] - \mathbb{E}_{s\sim d_0}[V_M^\pi(s)] \quad \Delta.$$

$$= \frac{1}{1-\gamma}\, \mathbb{E}_{s\sim d^\pi,\, a\sim\pi}\Big[f(s) - \underbrace{(r - \gamma f(s'))}_{\mathbb{E}[\cdots|s]} \Big]$$

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

$\mathbb{E}[\cdots|s] = T^\pi f$.

Let $f = V_{\hat{M}}^\pi$.

$$\mathbb{E}_{d_0}[V_{\hat{M}}^\pi - V_M^\pi] = \frac{1}{1-\gamma}\, \mathbb{E}_{\substack{s\sim d^\pi \\ a\sim\pi}}\Big[V_{\hat{M}}^\pi(s) - (T_M^\pi V_{\hat{M}}^\pi)(s) \Big]$$

$$(T_{\hat{M}}^\pi V_{\hat{M}}^\pi)(s)$$

$$= R(s,\pi) + \gamma \langle P(\cdot|s,\pi), V_{\hat{M}}^\pi \rangle - \hat{R}(s,\pi) - \gamma \langle \hat{P}(\cdot|s,\pi), V_{\hat{M}}^\pi \rangle.$$

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

Proof: $\frac{1}{1-\gamma}\,\mathbb{E}_{s,a,r,s'\sim d^\pi,\ a\sim\pi,\ r,s'}\left[f(s)-r-\gamma f(s')\right]$.

$= \sum_{t=1}^{\infty}\gamma^{t-1}\cdot\ \mathbb{E}_{s\sim d_t^\pi,\ a\sim\pi}\left[f(s)-r-\gamma f(s')\right]$.

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

$\mathbb{E}_{s\sim d_1^\pi,\ a\sim\pi}\left[f(s)-r-\gamma f(s')\right]$ +

$+\gamma\,\mathbb{E}_{s\sim d_2^\pi,\ a\sim\pi}\left[f(s)-r-\gamma f(s')\right]$

$+\gamma^2\,\mathbb{E}_{d_3^\pi,\ \ }\left[f(s)-r-\gamma f(s')\right]$.

$\vdots$

$J_M(\pi) = \mathbb{E}_{s\sim d_0}\left[V_M^\pi\right]$

$P=Q. \iff \forall f,\ \mathbb{E}_P[f]=\mathbb{E}_Q[f]$