Tabular Analysis of model-based RL

Setting: for each $(s,a) \in S \times A$.

$D_{s,a}$ = a bag of n i.i.d. (r, s'), where $\begin{cases} r \sim R(s,a) \\ s' \sim P(\cdot|s,a) \end{cases}$ $r \in [0, R_{max}]$.

Algorithm: Build empirical MDP: $\forall s, a$.

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

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

$e_{s'}$: $[0, 0, \cdots, 0, \underset{\uparrow}{1}, 0, \cdots, 0]$ (s'-th coordinate.)

$\longrightarrow$ empirical freq.

$$\hat{P}(s'|s,a) = \frac{\sum_{\tilde{s}} \mathbb{I}[\tilde{s} = s']}{n}.$$

Let $\hat{\pi}$ be the opt. policy in $\hat{M} = (S, A, \hat{P}, \hat{R}, \gamma)$, i.e. $\hat{\pi} = \pi^*_{\hat{M}}$

Want to bound $\| V^*_M - V^{\hat{\pi}}_M \|_\infty$.

certainty-equivalence

Analysis: { Step 1: establish $\hat{M} \approx M$ ⟵

Step 2: bound $V^*_M - V^{\hat{\pi}}_M$ using the error

How to measure the error of $\hat{M}$ (w.r.t. M)?

$$\varepsilon_R = \max_{s,a} |R(s,a) - \hat{R}(s,a)|$$

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

$\|u\|_1 = \sum_i |u_i|$.

Analysis: Start w/ step 2:

$$\forall s,\quad V_M^*(s) - V_M^{\hat{\pi}}(s) = V_M^{\pi_M^*}(s) - V_M^{\pi_{\hat{M}}^*}(s)$$

$$\leq V_M^{\pi_M^*}(s) - V_{\hat{M}}^{\pi_M^*}(s) + V_{\hat{M}}^{\pi_M^*}(s) - V_M^{\pi_{\hat{M}}^*}(s)$$

≥ 0 b/c $\pi_{\hat{M}}^*$ is opt. in $\hat{M}$

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

Simulation Lemma $\forall \pi: S \to A$. $\leq \frac{R_{max}}{1-\gamma}$.

$$\| V_M^\pi - V_{\hat{M}}^\pi \|_\infty \leq \frac{\varepsilon_R + \gamma \varepsilon_P V_{max}/2}{1-\gamma}$$

Proof: $\forall s,\quad |V_M^\pi(s) - V_{\hat{M}}^\pi(s)|$

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

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

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

$$(A) \leq \| P(s,\pi) - \hat{P}(s,\pi) \|_1 \cdot \| V_M^\pi \|_\infty$$

Slightly tighter: $\langle P(s,\pi), \vec{1} \rangle = 1$.

$$(A) = \left| \langle P(s,\pi) - \hat{P}(s,\pi), V_M^\pi - \frac{V_{max}}{2} \vec{1} \rangle \right|$$

$$\leq \| P(s,\pi) - \hat{P}(s,\pi) \|_1 \cdot \| V_M^\pi - \frac{V_{max}}{2} \vec{1} \|_\infty$$

$$\leq \varepsilon_P \cdot \frac{V_{max}}{2}. \qquad \left[-\frac{V_{max}}{2}, \frac{V_{max}}{2} \right]$$

Hölder's ineq:
$\forall p, q$, s.t. $\frac{1}{p} + \frac{1}{q} = 1$
$|\langle u, v \rangle| \leq \|u\|_p \cdot \|v$
Special cases:
(1) $p = q = 2$.
(2) $p = 1, q = \infty$.

$$(B) \leq \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty.$$

Dual norm: $\|\cdot\|, \|\cdot\|_*$
$\|x\|_* = \sup_{\|y\| \leq 1} y^T x$. ←

Put together: $\| V_M^\pi - V_{\hat{M}}^\pi \|_\infty \leq \varepsilon_R + \gamma \cdot \varepsilon_P \cdot \frac{V_{max}}{2} + \gamma \cdot \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty$

$$\Rightarrow \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty \leq \frac{\varepsilon_R + \gamma \varepsilon_P V_{max}/2}{1-\gamma}.$$

$$\| V_M^* - V_{\hat{M}}^{\hat{\pi}} \|_\infty \leq 2 \max_\pi \| V_M^\pi - V_{\hat{M}}^\pi \| \leq 2 \cdot (\,)$$

Step 1: $M \approx \hat{M}$ (i.e. upper bound ε_R & ε_P).

assume stoch. rewards bounded in $[0, R_{max}]$.

Fix any s,a. w.p. $\geq 1-\delta$,

$$|\hat{R}(s,a) - R(s,a)| \leq R_{max} \cdot \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$$

Fix any s,a

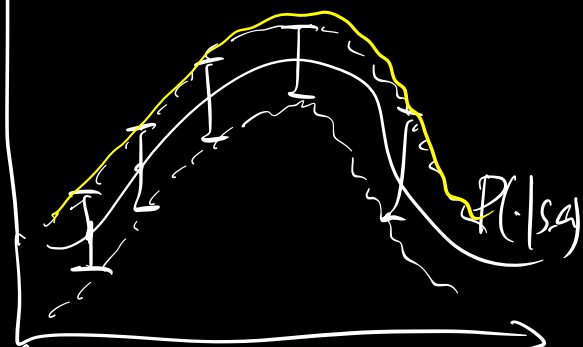
Naive analysis:

$$\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 = \sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)|$$

$$\leq |S| \cdot \max_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)|$$

$$\frac{1}{n} \sum_{\tilde{s} \in D_{s,a}} \mathbb{I}[\tilde{s} = s'] \quad \Big| \quad \approx |S| \, O\left(\frac{1}{\sqrt{n}}\right)$$

Lemma: Let p be a multinomial dist. over S. Let $\hat{p}$ be emp. dist. estimated from n samples i.i.d from p. Then $\|p - \hat{p}\|_1 \leq 2 \cdot \sqrt{\frac{1}{2n} \ln \frac{2 \cdot 2^{|S|}}{\delta}}$, w.p. $\geq 1-\delta$.



Proof: For any vector $v \in \mathbb{R}^s$.

$$\|v\|_1 = \max_{u \in \{-1,1\}^s} \underbrace{u^T v}$$

$$\leq \|u\|_\infty \cdot \|v\|_1 = \|v\|_1$$

Proof of LHS=RHS:

$\|v\|_1 \leq$ RHS.

RHS $\leq \|v\|_1$ (Hölder).

Therefore: $\|p - \hat{p}\|_1$

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

$$= \max_{u \in \{-1,1\}^s} \left(u^T p - u^T \hat{p}\right).$$

Since $\hat{p} = \frac{1}{n} \sum_{s \in D} e_s$, where $e_s = [0, \cdots, 0, 1, 0, \cdots, 0]^T$, D is the i.i.d. samples drawn $\sim p$.

$$u^T p - u^T \hat{p} = u^T p - \underbrace{\frac{1}{n} \sum_{s \in D} u^T e_s}$$

Now, $u^T e_s$ is iid (scalar valued) r.v. with mean $u^T p$.

Also, $|u^T e_s| \leq \|u\|_\infty \cdot \|e_s\|_1 = 1$.

$\hookrightarrow \in \{-1, 1\}$.

$\Rightarrow$ Fixing u, Hoeffding's: w.p. $\geq 1-\delta$.

$$\left| u^T p - \frac{1}{n} \sum_{s \in D} u^T e_s \right| \leq 2 \cdot \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}.$$

By union bound: w.p. $\geq 1-\delta$,

$$\|p - \hat{p}\|_1 = \max_{u \in \{-1,1\}^S} \left| u^T p - u^T \hat{p} \right| \leq 2 \cdot \sqrt{\frac{1}{2n} \ln \frac{2 \cdot 2^{|S|}}{\delta}}.$$

$$\approx \sqrt{\frac{|S|}{n}}$$

Finally, union bound over $|S \times A|$ reward estimations & $|S \times A|$ transition estimations.

$$\varepsilon_R \leq R_{max} \cdot \sqrt{\frac{1}{2n} \ln \frac{2 \cdot 2 \cdot |S \times A|}{\delta}}$$

$$\varepsilon_p \leq 2 \cdot \sqrt{\frac{1}{2n} \cdot \ln \frac{2 \cdot 2 \cdot |S \times A| \cdot 2^{|S|}}{\delta}}.$$

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

$$\|V_M^* - V_M^{\hat{\pi}}\|_\infty \leq \boxed{\tilde{O}\left(\frac{\sqrt{S} \cdot V_{max}}{\sqrt{n}\,(1-\gamma)}\right)}$$

ignoring logarithms

Alternative Analysis

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

$$\|V_M^* - V_M^{\hat{\pi}}\|_\infty \leq \tilde{O}\left(\frac{V_{max}}{\sqrt{n}(1-\gamma)^2}\right).$$

In the previous analysis: $\forall \pi$, $V_{\hat{M}}^{\pi} \approx V_M^{\pi}$.

We will use a different way to establish the near-opt of $\hat{\pi}$.

Recall: $\forall f \in \mathbb{R}^{S\times A}$, $\|V_M^* - V_M^{\pi_f}\|_\infty \leq \frac{2\cdot\|f - Q_M^*\|_\infty}{1-\gamma}$.

If we want to let $\hat{\pi} = \pi_f$, what is f?

Answer: $f = Q_{\hat{M}}^*$. $\left(\hat{\pi} = \pi_{Q_{\hat{M}}^*}\right)$.

$$\Rightarrow \|V_M^* - V_M^{\hat{\pi}}\|_\infty \leq \frac{2}{1-\gamma}\cdot\|\underbrace{Q_{\hat{M}}^* - Q_M^*}_{\Delta}\|_\infty.$$

be careful: $Q_{\hat{M}}^* - Q_M^* \neq Q_{\hat{M}}^{\pi} - Q_M^{\pi}$ for the same π.

($Q_{\hat{M}}^*$: $\pi_{\hat{M}}^*$, Q_M^*: π_M^*)

$$\|Q_{\hat{M}}^* - Q_M^*\|_\infty = \|\underline{Q_{\hat{M}}^*} - \underline{T_{\hat{M}} Q_M^*} + T_{\hat{M}} Q_M^* - Q_M^*\|_\infty$$

$$\le \|T_{\hat{M}} Q^*_{\hat{M}} - T_{\hat{M}} Q^*_M\|_\infty + \|T_{\hat{M}} Q^*_M - Q^*_M\|_\infty.$$

$$\le \gamma \cdot \|Q^*_{\hat{M}} - Q^*_M\|_\infty + \|T_{\hat{M}} Q^*_M - Q^*_M\|_\infty.$$

$\forall\, s, a.$

$$\left|(T_{\hat{M}} Q^*_M)(s,a) - Q^*_M(s,a)\right| \qquad V^*_M(s')$$

$$= \left| \hat{R}(s,a) + \gamma \mathbb{E}_{s' \sim \hat{P}(\cdot|s,a)} \left[\max_{a'} Q^*_M(s',a') \right] - Q^*_M(s,a) \right|$$

$$= \left| \frac{1}{n} \sum_{r \in D_{s,a}} r + \gamma \left\langle \frac{1}{n} \sum_{s' \in D_{s,a}} e_{s'}, V^*_M(\cdot) \right\rangle - Q^*_M(s,a) \right|$$

$$= \left| \frac{1}{n} \sum_{(r,s') \in D_{s,a}} \left(r + \gamma \langle e_{s'}, V^*_M \rangle \right) - Q^*_M(s,a) \right|.$$

$$[0, 0, \cdots, 0, 1, 0, 0, \cdots 0]$$

s'-th coordinate

$$= \left| \frac{1}{n} \sum_{(r,s') \in D_{s,a}} \left(r + \gamma V^*_M(s') \right) - Q^*_M(s,a) \right| \qquad (*).$$

empirical Bellman update

Notice that since (r, s') in $D_{s,a}$ are iid.

$\Rightarrow$ $\underbrace{r + \gamma V_M^*(s')}_{\text{scalar.} \in [0, \frac{R_{max}}{1-\gamma}]}$ are also iid for $(r, s') \in D_{s,a}$.

$r \le R_{max}$. $V_M^* \le \frac{R_{max}}{1-\gamma}$.

Fix s, a. $\mathbb{E}_{r,s'}[r + \gamma V_M^*(s')]$ $\max_{a'} Q_M^*(s', a') = V_M^*(s')$

$= R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_M^*(s')]$.

$= (T_M Q_M^*)(s,a) = Q_M^*(s,a)$.

By Hoeffding's. w.p $\ge 1-\delta$.

$(*) \le \frac{R_{max}}{1-\gamma} \cdot \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$.

By union bound, w.p $\ge 1-\delta$.

$\underline{\| T_{\hat{M}} Q_M^* - Q_M^* \|_\infty \le V_{max} \sqrt{\frac{1}{2n} \ln \frac{2|S \times A|}{\delta}}}$

$\Rightarrow \| V_M^* - V_M^{\hat{\pi}} \|_\infty \le \tilde{O}\left(\frac{V_{max}}{\sqrt{n}(1-\gamma)^2} \right)$

Remark 1. $\forall f \in \mathbb{R}^{S\times A}$. $\mathbb{E}[\cdots \mid s,a] = (\mathcal{T}f)(s,a)$.

$$(\hat{\mathcal{T}}_n f)(s,a) = \frac{1}{n} \sum_{(r,s')\in D_{s,a}} \left(r + \gamma \max_{a'} f(s',a')\right)$$

emp. Bellman update of f.

Previous analysis

$\hat{\mathcal{T}}_n \approx \mathcal{T}_n$. in the sense that.

$\forall f$ (w/ bounded range). $\hat{\mathcal{T}}_n f \approx \mathcal{T}_n f$

This analysis: $\hat{\mathcal{T}}_n f \approx \mathcal{T}_n f$ only holds for $f = Q_n^*$.

Remark 2: this analysis uses.

$$\|Q_n^* - Q_{\hat{n}}^*\|_\infty \le \frac{\|\mathcal{T}_n Q_n^* - \hat{\mathcal{T}}_n Q_n^*\|_\infty}{1-\gamma}$$

But the following also holds:

$$\|Q_n^* - Q_{\hat{n}}^*\|_\infty \le \frac{\|\hat{\mathcal{T}}_{\hat{n}} Q_{\hat{n}}^* - \mathcal{T}_n Q_{\hat{n}}^*\|_\infty}{1-\gamma}.$$

↓

$$\left| \frac{1}{n} \sum_{(r,s') \in D_{s,a}} (r + \gamma V_M^*(s')) - \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V_M^*(s')] \right) \right|$$

Hoeffding's doesn't apply

$V_M^*(s')$ — Random

For fixed $f \in \mathbb{R}^S$.

$$\left| \frac{1}{n} \sum_{(r,s') \in D_{s,a}} (r + \gamma f(s')) - \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [f(s')] \right) \right|$$

Hoeffding's applies!