2.2 Sublinear Sample Complexity

Previously, we are able to estimate value of any policy. Now we focus on the estimate of $\hat{Q}^*$.

#samples from generative model $= |S||A|N$

Lemma 2.5 [component-wise bounds]

$$Q^* - \hat{Q}^* \le \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^* \quad (1)$$

$$Q^* - \hat{Q}^* \ge \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P - \hat{P}) V^* \quad (2)$$

Pf: (1): π^* is optimal for M, $\hat{\pi}$ is optimal for $\hat{M}$

$$\begin{aligned} Q^* - \hat{Q}^* &= Q^{\pi^*} - \hat{Q}^{\hat{\pi}} \\ &\le Q^{\pi^*} - \hat{Q}^{\pi^*} \\ \text{(Lemma 2.2)} &= \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^{\pi^*} \end{aligned}$$

(2)

$$\begin{aligned} Q^* - \hat{Q}^* &= Q^{\pi^*} - \hat{Q}^{\hat{\pi}} \\ &= \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P^{\pi^*} - \hat{P}^{\hat{\pi}}) Q^* \\ &\ge \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P^{\pi^*} - \hat{P}^{\pi^*}) Q^* \\ &= \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P - \hat{P}) V^* \end{aligned}$$

where $\hat{P}^{\hat{\pi}} Q^* \le \hat{P}^{\pi^*} Q^*$ is because $\pi^*(s) = \arg\max_{a \in A} Q^*(s,a)$

Proposition 2.4 [Crude Value Bounds]

Let $\delta \ge 0$, with prob. $\ge 1 - \delta$,

$$\|Q^* - \hat{Q}^*\|_\infty \le \Delta_{\delta, N}$$

$$\|Q^* - \hat{Q}^{\pi^*}\|_\infty \le \Delta_{\delta, N},$$

where

$$\Delta_{\delta,N} := \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2\log(2|S||A|/\delta)}{N}}$$

Pf: Due to Lemma 2.2 and Lemma 2.3

$$\|Q^* - \hat{Q}^{\pi^*}\|_\infty \le \frac{\gamma}{1-\gamma}\|(P-\hat{P})V^*\|_\infty \quad (3)$$

Due to Lemma 2.5 and Lemma 2.3

$$\|Q^* - \hat{Q}^*\|_\infty \le \frac{\gamma}{1-\gamma}\|(P-\hat{P})V^*\|_\infty \quad (4)$$

By applying Hoeffding's ineq.

$$\|(P-\hat{P})V^*\|_\infty = \max_{s,a}\left|E_{s'\sim P(s,a)}[V^*(s')] - E_{s'\sim\hat{P}(s,a)}[V^*(s')]\right|$$
$$\le \frac{1}{1-\gamma}\sqrt{\frac{2\log(2|S||A|/\delta)}{N}},$$

which holds with prob. $\ge 1-\delta$. □

Addition:

Hoeffding's ineq. : $P(|S_n - E[S_n]| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$

where $a_i \le X_i \le b_i$, $S_n = \sum_{i=1}^n X_i$.

Since $E_{s'\sim\hat{P}(s,a)}[V^*(s')] = \sum_{i=1}^N V^*(s_i)/N$

and $0 \le V^*(s_i)/N \le \frac{1}{(1-\gamma)N}$, we have

$$P\left(\left|E_{s'\sim P(s,a)}[V^*(s^*)] - E_{s'\sim\hat{P}(s,a)}[V^*(s')]\right| \ge \frac{1}{1-\gamma}\sqrt{\frac{2\log(2|S||A|/\delta)}{N}}\right)$$

$$\le 2\exp\left\{\frac{-2\cdot\frac{1}{(1-\gamma)^2}\cdot\frac{2\log(2|S||A|/\delta)}{N}}{\frac{1}{(1-\gamma)^2}N^2}\right\}$$

$$= \exp\{-4N\}\frac{1}{|S||A|}\delta$$

$$\le \frac{1}{|S||A|}\delta$$