2.1 a naive model-based approach

Central question:

Do we require an accuracy model of the world to find a near optimal policy?

A naive model to learn P: after sampling N times,

let $\hat{P}(s'|s,a) = \dfrac{\text{count}(s',s,a)}{N}$

here we view $\hat{P}$ as a matrix of size $|S||A| \times |S|$

Expectation: $O(|S|^2|A|)$ obeservation is enough for an accurate model.

Proposition 2.1 Assume $\varepsilon \in (0, \frac{1}{1-\gamma})$, $\exists\ c > 0$ s.t.

samples from generative model

$= |S||A| N \geq \dfrac{4c^2}{(1-\gamma)^4} \dfrac{|S|^2|A|\log(|S||A|/\delta)}{\varepsilon^2}$

where (s,a) is sampled uniformly. and with prob. $> 1-\delta$

we have

① (Model accuracy)

$$\max_{s,a} \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2 \varepsilon$$

② (Uniform value accuracy)

$$\|Q^\pi - \hat{Q}^\pi\|_\infty \leq \frac{\varepsilon}{2} \quad \text{for all } \pi$$

③ (Near optimal planning) Suppose $\hat{\pi}$ is optimal w.r.t. $\hat{M}$

$$\|\hat{Q}^* - Q^*\|_\infty \leq \frac{\varepsilon}{2}, \quad \|Q^{\hat{\pi}} - Q^*\|_\infty \leq \varepsilon$$

To show this, we need following lemmas.

> **Lemma 2.2** [Simulation lemma] For all π:
> $$Q^\pi - \hat{Q}^\pi = \gamma (I - \gamma \hat{P}^\pi)^{-1} (P - \hat{P}) V^\pi.$$

Pf:
$$\begin{aligned}
Q^\pi - \hat{Q}^\pi &= Q^\pi - (I - \gamma \hat{P}^\pi)^{-1} r \\
&= (I - \gamma \hat{P}^\pi)^{-1} \left((I - \gamma \hat{P}^\pi) Q^\pi - r \right) \\
&= (I - \gamma \hat{P}^\pi)^{-1} \left((I - \gamma \hat{P}^\pi) - (I - \gamma P^\pi) \right) Q^\pi \\
&= \gamma (I - \gamma \hat{P}^\pi)^{-1} (P^\pi - \hat{P}^\pi) Q^\pi \\
&= \gamma (I - \gamma \hat{P}^\pi)^{-1} (P - \hat{P}) V^\pi \qquad \square
\end{aligned}$$

> **Lemma 2.3** For any policy π, MDP M and $v \in \mathbb{R}^{|S||A|}$
> $$\|(I - \gamma P^\pi)^{-1} v\|_\infty \le \frac{1}{1-\gamma} \|v\|_\infty$$

Pf: $v = (I - \gamma P^\pi)(I - \gamma P^\pi)^{-1} v =: (I - \gamma P^\pi) w$

$$\begin{aligned}
\Rightarrow \|v\|_\infty &= \|(I - \gamma P^\pi) w\|_\infty \\
&\ge \|w\|_\infty - \gamma \|P^\pi w\|_\infty \\
&\ge \|w\|_\infty - \gamma \|w\|_\infty \\
&= (1-\gamma) \|w\|_\infty
\end{aligned}$$

i.e. $\|(I - \gamma P^\pi)^{-1} v\| \le \frac{1}{1-\gamma} \|v\|_\infty$ $\square$

> **Lemma A.8** [concentration for discrete distributions]
>
> Let z be r.v. of $\{1, \cdots, d\}$, distributed according to q, where $\vec{q} = [P_r(z=j)]_{j=1}^d$. Assume we have N i.i.d. samples and that our empirical estimate is $[\hat{q}]_j = \sum_{i=1}^N \mathbb{1}\{z_i = j\} / N$, we have $\forall \varepsilon > 0$:
> $$P_r\left(\|\hat{q} - q\|_2 \ge \frac{1}{\sqrt{N}} + \varepsilon\right) \le e^{-N\varepsilon^2},$$
> which implies:

$$\Pr\left(\|\hat{q}-\vec{q}\|_1 \geq \sqrt{d}\left(\frac{1}{\sqrt{N}}+\varepsilon\right)\right) \leq e^{-N\varepsilon^2}.$$

this proof is ignored

Pf of Proposition 2.1 :

with l_1 norm in lemma A.8, for fixed s, a, with prob. $\geq 1-\delta$, we have

$$\|P(\cdot|s,a)-\hat{P}(\cdot|s,a)\|_1 \leq c\sqrt{\frac{|S|\log(1/\delta)}{N}} \qquad (*)$$

where N is the number of samples used to estimate $\hat{P}(\cdot|s,a)$. Let $|S||A|\delta = e^{-N\varepsilon^2} \Rightarrow \varepsilon = \sqrt{\frac{\log(|S||A|/\delta)}{N}}$, $d=|S|$

and let c satisfy $c\sqrt{\frac{|S|\log(|S||A|/\delta)}{N}} \geq \sqrt{|S|}\left(\frac{1}{\sqrt{N}}+\varepsilon\right)$

① $\|P(\cdot|s,a)-\hat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2\varepsilon$

since $N \geq \frac{4c^2}{(1-\gamma)^4}\frac{|S|\log(|S||A|/\delta)}{\varepsilon^2}$, by $(*)$ we have

$\|P(\cdot|s,a)-\hat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2\varepsilon/2$ with prob. $\geq 1-\delta$

② $\|Q^\pi - \hat{Q}^\pi\|_\infty \leq \frac{\varepsilon}{2}$

By Lemma 2.2:

$$\|Q^\pi - \hat{Q}^\pi\|_\infty = \|\gamma(I-\gamma P^\pi)^{-1}(P-\hat{P})V^\pi\|_\infty$$

Lemma 2.3 $\leq \frac{\gamma}{1-\gamma}\|(P-\hat{P})V^\pi\|_\infty$

Hölder ineq. $\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot|s,a)-\hat{P}(\cdot|s,a)\|_1\right)\|V^\pi\|_\infty$

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot|s,a)-\hat{P}(\cdot|s,a)\|_1$$

$$\leq \gamma\varepsilon/2 \leq \varepsilon/2$$

③ $\|\hat{Q}^* - Q^*\|_\infty \leq \frac{\varepsilon}{2}$, $\|Q^{\hat{\pi}} - Q^{\pi^*}\|_\infty \leq \varepsilon$

observe that $|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x)-g(x)|$

$$\Rightarrow |\hat{Q}^*(s,a) - Q^*(s,a)| = |\sup_\pi \hat{Q}^\pi(s,a) - \sup_\pi Q^\pi(s,a)|$$

$$\leq \sup_\pi |\hat{Q}^\pi(s,a) - Q^\pi(s,a)|$$

$$
\begin{aligned}
\|Q^{\hat\pi} - Q^{\pi^*}\|_\infty &\le \overbrace{\|Q^{\hat\pi} - \hat Q^{*}\|_\infty}^{\le \frac{\varepsilon}{2}} + \|\hat Q^{*} - Q^{\pi^*}\|_\infty \\
&= \|Q^{\hat\pi} - \hat Q^{\hat\pi}\|_\infty + \|\hat Q^{*} - Q^{*}\|_\infty \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \square
\end{aligned}
$$