RECALL HOEFFDING:

$$\Pr\left[\left|S_n - \mathbb{E}(S_n)\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

$$\Pr\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}(S_n)}{n}\right| \geq \frac{t}{n}\right] \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

Fix s, a, s'

$$\Pr\left[\left|\hat{P}(s'|s,a) - P(s'|s,a)\right| \geq \frac{t}{n}\right] \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

$$\text{Probability}\left(\exists s, a \left|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\right| \geq t/n\right) \leq 2|S||A|\exp\left(\frac{-2t^2}{n}\right)$$

$\exists s'$

We want : $2|S|^2|A|\exp\left(\frac{-2t^2}{n}\right) \leq \delta$.

If $\forall$ policies π we want $\|Q^\pi - Q\|_\infty \leq \varepsilon/2$

then we want

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

$$\therefore \quad \frac{t}{n} = \frac{(1-\gamma)^2 \varepsilon}{2} \qquad \therefore \quad t = \frac{(1-\gamma)^2 \varepsilon n}{2}$$

$$\therefore \quad 2|S|^2|A| \exp\left(-2\frac{(1-\gamma)^4 \varepsilon^2}{4} n\right) \leq \delta$$

$\therefore$

$$\exp\left(\frac{(1-\gamma)^4 \varepsilon^2 n}{2}\right) \geq \frac{2|S|^2|A|}{\delta}$$

$$n \geq \frac{2}{(1-\gamma)^4 \varepsilon^2} \log\left(\frac{2|S|^2|A|}{\delta}\right)$$

$$\therefore \quad \underline{\#\text{samples}} = \frac{2|S|^2|A|}{(1-\gamma)^4 \varepsilon^2} \log\left(\frac{2|S|^2|A|}{\delta}\right)$$

$$= \tilde{O}\left(\frac{N}{(1-\gamma)^4 \varepsilon^2}\right)$$

Can one improve this?

Turns out we can:

We do not need $\| Q^{\pi} - \hat{Q}^{\pi} \|_{\infty}$ for all policies. Need only for Q^* !!

More refined:

Lemma: Let $\delta \geq 0$. with prob $\geq 1-\delta$

$$\| Q^* - \hat{Q}^{\pi^*} \|_\infty \leq \frac{\gamma}{1-\gamma} \sqrt{\frac{2\log(2|S||A|/\delta)}{N}}$$

Pf. $\| Q^* - \hat{Q}^{\pi^*} \|_\infty = \gamma \| P^{\pi^*} Q^* - \hat{P}^{\pi^*} \hat{Q}^{\pi^*} \|_\infty$

$$\leq \gamma \| P^{\pi^*} Q^* - \hat{P}^{\pi^*} Q^* \|_\infty + \gamma \| \hat{P}^{\pi^*} Q^* - \hat{P}^{\pi^*} \hat{Q}^{\pi^*} \|_\infty$$

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

$$\leq \gamma \| P V^* - \hat{P} V^* \|_\infty + \gamma \| Q^* - \hat{Q}^{\pi^*} \|_\infty$$

$$\therefore \| Q^* - \hat{Q}^{\pi^*} \| \leq \frac{\gamma}{1-\gamma} \| P V^* - \hat{P} V^* \|_\infty .$$

$$\| P V^* - \hat{P} V^* \|_\infty = \max_{s,a} \left| \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^*(s')] - \mathbb{E}_{s' \sim \hat{P}(\cdot|s,a)} [V^*(s')] \right|$$

If this is $\leq \sqrt{\frac{2\log(2|S||A|/\delta)}{N}}$ we are done!

Use Hoeffdings inequality.

Consider the vector

$$V = \boxed{V^*(s_1) \mid V^*(s_2) \mid \; - \; - \; - \quad - \; -}$$

Pick $s' \sim P(\cdot \mid s,a)$ and consider the random variable $x = V(s')$.

$$\mathbb{E}(x) = \sum_{s' \sim P(\cdot \mid s,a)} P(s' \mid s,a)\, V(s')$$

Instead pick $s_1 \sim P(\cdot \mid s,a)$, $s_2 \sim P(\cdot \mid s,a)$, ...
and consider $x_1 = V(s_1)$, $x_2 = V(s_2)$, ...

$$\mathbb{E}(x_1) = \mathbb{E}(x_2) = \cdots = \sum_{s' \sim P(\cdot \mid s,a)} P(s' \mid s,a)\, V(s')$$

Take $X = \dfrac{x_1 + \cdots + x_N}{N}$ $\qquad \mathbb{E}(x) = \displaystyle\sum_{s' \sim P(\cdot \mid s,a)} P(s' \mid s,a)\, V(s')$

Clearly each $x_i \in [0,1]$

$$\therefore \Pr\left[\left| x - \mathbb{E}(x) \right| > \frac{t}{N} \right] \leq 2\exp\left(\frac{-2t^2}{N} \right)$$

Prob that for some (s,a) the above fails is at most

$$2|S|\,|A| \exp\left(-\frac{2\cancel{t}^2}{N}\right)$$

want this to be $\leq \delta$.

$$\therefore \quad \exp\left(\frac{2t^2}{\cancel{N}}\right) \geq \frac{2|S|\,|A|}{\delta}$$

$$t^2 \geq \frac{N}{2} \ln\left(\frac{2|S|\,|A|}{\delta}\right)$$

$$\therefore \quad t \geq \sqrt{\frac{N \ln\left(\frac{2|S|\,|A|}{\delta}\right)}{2}}$$

$$\|(P-\hat{P})V^*\|_\infty \leq \frac{t}{N} = \sqrt{\frac{\ln\left(\frac{2|S|\,|A|}{\delta}\right)}{2N}}$$ with

prob $(1-\delta)$

Now set:

$$\frac{\gamma}{1-\gamma} \sqrt{\frac{\log(2|S|\,|A|/\delta)}{2N}} \leq \varepsilon$$

$$\boxed{N \geq \frac{\gamma^2}{(1-\gamma)^2 \varepsilon^2} \frac{1}{2} \log\left(2|S||A|/\delta\right)}$$

The factor is $\frac{1}{(1-\gamma)^2 \varepsilon^2}$

$\therefore$ # samples: $\frac{|S||A|}{(1-\gamma)^2 \varepsilon^2} \, 2 \log\left(\frac{2|S||A|}{\delta}\right)$

But this requires $V^*(s')$ - we have no access to that

<u>Thm:</u> (Azar, Munos, Kappen)

With prob $(1-\delta)$, an order of $O\left(\frac{N \log(N/\delta)}{(1-\gamma)^3 \varepsilon^2}\right)$

Samples suffice to find ε-optimal estimate of action value, and to find an ε-optimal policy.

- There is a (almost) matching upper bound

<u>Thm:</u>

$$\|Q^* - \hat{Q}^*\|_\infty \leq \gamma \sqrt{\frac{c \log(c|S||A|/\delta)}{N}} + \frac{c\gamma}{(1-\gamma)^2} \frac{\log(c|S||A|/\delta)}{N}$$

= <u>Strategic Exploration:</u>

- Don't have access to transitions at each state.
- Can execute trajectories in the MDP.

- Agent has to engage in exploration - to reach new states where not enough samples are

Seen!

The algorithm maintains an estimate of transition prob $P(s'|s,a)$ $\forall s'$, nbrs of s.

- It also estimates reward.
- If s is visited enough, declare it as known.
- Learning is complete when all states are known.
- In a known state - take the optimal action

Input: parameter m; ε;

1. K (known states) $= \phi$. $\forall s, a, s'$, $n(s,a) = n(s,a,s') = 0$
   $R(s,a) = 0$

2- Observe initial state s_0, π_0 be an initial policy.

3 **for** rounds $t = 0, 1, \ldots$

4 **if** a state has become known (i.e. $n(s,a) \geq m \; \forall a$) **then**

5 update $K = K \cup \{s\}$.

6 Let $\hat{M}$ have $\hat{P}(s' | s,a) = \dfrac{n(s,a,s')}{n(s,a)}$, $\hat{r}(s,a) = \dfrac{R(s,a)}{n(s,a)}$

7 Let $\hat{M}_K$ be induced MDP.
$\pi_t = \pi^*(\hat{M}_K)$ - optimal policy in $\hat{M}_K$

8 **else**

9 if $t \geq 1$ $\pi_t = \pi_{t-1}$

10 **endif**

11 If $s_t \in K$, $a_t = \pi_t(s_t)$, else $a_t = \arg\min_a n(s,a)$

12 Get reward r_t & observe s_{t+1}

13 **if** $s_t \notin K$ **then**

14 update $n(s_t, a_t) += 1$; $R(s_t, a_t) += r_t$;
$n(s_t, a_t, s_{t+1}) += 1$;

15 **endif**

16 **endfor**

INDUCED MDP: M is given;

Assumes $K \subseteq S$;

$\forall s \in K$, $P_{M_K}(s' | s, a) = P_M(s' | s, a)$ &

$r_{M_K}(s, a) = r_M(s, a)$

$\forall s \notin K$, $P_{M_K}(s' | s, a) = \mathbb{1}(s' = s)$ and

$r_{M_K}(z | s, a) = \mathbb{1}(z = 1)$