Online RL | Tabular : $M = (S, A, P, R, \gamma, d_0)$.

(assume all traj. from $s \sim d_0$ terminates in H steps)

For $t = 1, 2, 3, \cdots, T$

- Learner collects trajectory using π_t.
- $(s_1, a_1, r_1, \cdots, s_H, a_H, r_H)$

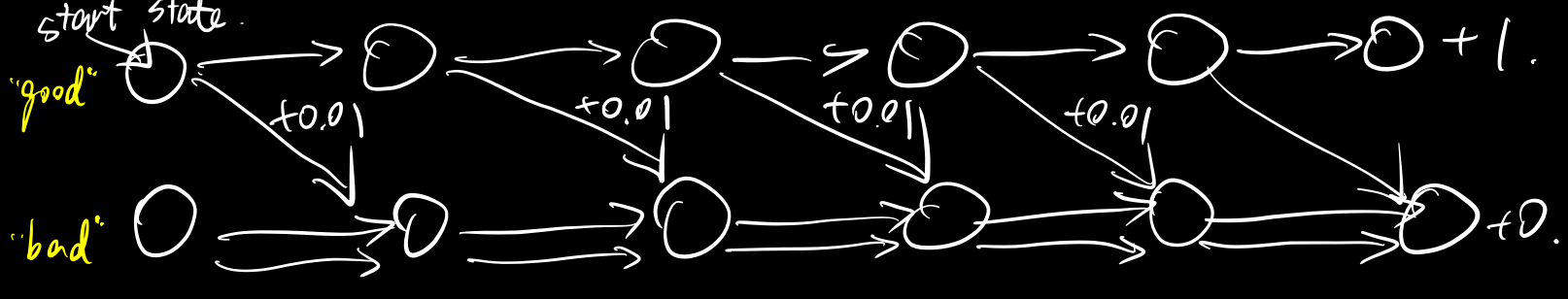
After T trajectories, output policy $\hat{\pi}$.

Guarantee w.p. $\geq 1-\delta$, $J(\pi^*) - J(\hat{\pi}) \leq \varepsilon \cdot V_{max}$.

w/ sample complexity $T = \text{poly}(|S|, |A|, \frac{1}{1-\gamma}, \frac{1}{\varepsilon}, \frac{1}{\delta})$.

Unif exploration is ineffective.

"Combination Lock" (not fundamental hardness).



H.

if explore w/ unif over actions.

(extend to ε-greedy, softmax).

Rmax At any point during exploration, maintain.

- $n(s,a)$: #times we see (s,a) in data.
- $n(s,a,s')$: #times we see (s,a,s') in data.

* stop accumulating data after $n(s,a) = m$. (artifact for clean concentration analysis).

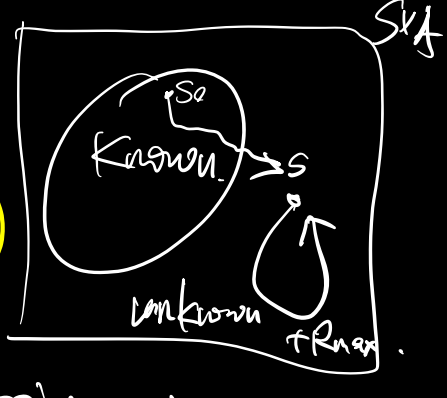
Rmax: · define $K := \{(s,a) : n(s,a) = m\}$. "Known".
↳ threshold.

· Build MDP $\hat{M}_K$:

$$\hat{P}_K(s'|s,a) = \begin{cases} n(s,a,s')/n(s,a), & \text{if } (s,a) \in K. \\ \mathbb{I}[s'=s]. & \text{if } (s,a) \notin K. \end{cases}$$

$$\hat{R}_K(s,a) = \begin{cases} R(s,a), & \text{if } (s,a) \in K. \\ R_{max}. & \text{if } (s,a) \notin K. \end{cases}$$

· next-exploration policy: $\pi^*_{\hat{M}_K}$.



"Optimism in face of uncertainty" (OFU)

Remark: alg needs to determine stopping criteria.
(ignored for now; will comment on later)

Define M_K:

$$P_K(s'|s,a) = \begin{cases} P(s'|s,a), & \text{if } (s,a) \in K. \\ \mathbb{I}[s'=s], & \text{o.w.} \end{cases}$$

$$R_K(s,a) = \begin{cases} R(s,a), & \text{if } (s,a) \in K. \\ R_{max}. & \text{o.w.} \end{cases}$$

	M	M_K	$\hat{M}_K$
$(s,a) \in K$	$=M$	$=M$	$\approx M$
$(s,a) \notin K$	$=M$	self-loop	self-loop

Define: $dis(M_K, \hat{M}_K) = \max_{s,a} \| P_K(\cdot|s,a) - \hat{P}_K(\cdot|s,a) \|_1$.

Lemma 1. For any fixed s,a, w/ $n(s,a) = m$.

w.p. $\geq 1-\delta$, $\| P_K(\cdot|s,a) - \hat{P}_K(\cdot|s,a) \|_1 \leq \sqrt{\frac{2}{m} \log \frac{2 \cdot 2^{|S|}}{\delta}}$.

Lemma 2. $\| V^*_{M_K} - V^*_{\hat{M}_K} \|_\infty \leq \frac{dis(M_K, \hat{M}_K) \cdot V_{max}}{2(1-\gamma)}$.

Lemma 3. $\forall \pi$, $\| V^\pi_{M_K} - V^\pi_{\hat{M}_K} \|_\infty \leq (\cdots)$.

Lemma 4. Consider M_K & M. $\forall \pi$.

$$\left| J_M(\pi) - J_{M_K}(\pi) \right| \leq V_{max} \cdot P_M[\text{escape}_K(\tau) | \pi]$$

where $\text{escape}_K(\tau) = \mathbb{I}[\exists h, (s_h, a_h) \notin K]$.

• $\mathbb{P}_M[\cdot \mid \pi]$ is the prob. when traj. τ is generated by running π in M.

Proof. $J_M(\pi) \overset{?}{\leq} J_{M_K}(\pi) + V_{max} \cdot \mathbb{P}_M[\cdots]$.

$$= \sum_\tau \mathbb{P}_M[\tau \mid \pi] \cdot R_M(\tau)$$

$$R_M(\tau) = \sum_{h=1}^{H} \gamma^{h-1} R(s_h, a_h).$$

$$= \sum \mathbb{P}_M[\tau \mid \pi] \cdot \left(R_M(pre_K(\tau)) + R_M(suf_K(\tau)) \right)$$

τ: $escape_K(\tau)=1$

$$+ \sum \mathbb{P}_M[\tau \mid \pi] \cdot R_M(\tau) \Big]$$

τ: $escape_K(\tau)=0$

$s_1, a_1, s_2, a_2, \cdots, s_h, a_h, \cdots, s_H, a_H$.

$pre_K(\tau)$ $\quad$ $suf_K(\tau)$.

(s_h, a_h): 1st (s,a) outside K.

$$\leq \sum_{esc=1} \mathbb{P}_M[\tau \mid \pi] \left(R_M(pre_K(\tau)) + V_{max} \right)$$

$$+ \sum_{esc=0} \mathbb{P}_M[\tau \mid \pi] R_M(\tau).$$

$$= \sum_{pre_K(\tau)} \mathbb{P}_M[pre_K(\tau) \mid \pi] \left(R_M(pre_K(\tau)) + V_{max} \right) + \sum_{esc=0} \mathbb{P}_M[\tau \mid \pi] R_M(\tau).$$

$$= \sum_{pre_K(\tau)} \mathbb{P}_{M_K}[pre_K(\tau) \mid \pi] \left(R_{M_K}(pre_K(\tau)) + V_{max} \right) + \sum_{esc=0} \mathbb{P}_{M_K}[\tau \mid \pi] R_{M_K}(\tau)$$

$$\mathbb{P}[\tau] = d_0(s_1) \cdot \pi(a_1 \mid s_1) \cdot P(s_2 \mid s_1, a_1) \cdots$$

$$\leq J_{M_K}(\pi) + \sum_{pre_K(\tau)} \mathbb{P}_{M_K}[pre_K(\tau) \mid \pi] \cdot V_{max}.$$

$$= J_{M_K}(\pi) + \left(\sum_{pre_K(\tau)} \mathbb{P}_M[pre_K(\tau) \mid \pi] \right) V_{max}$$

$$= J_{M_K}(\pi) + \mathbb{P}_M[escape_K(\tau) \mid \pi] \cdot V_{max} \quad \square$$

Remark: only uses the fact $M = M_K$ on K.
(didn't use the Rmax design).

Putting things together:

"optimal or explore": either near-optimal (in M) or explore (s.a. $\notin K$).

$\pi^*_{\hat{M}_K}$

$$\varepsilon \cdot V_{max} < J_M(\pi^*_M) - J_M(\pi^*_{\hat{M}_K}) \le J_{M_K}(\pi^*_M) - J_M(\pi^*_{\hat{M}_K})$$

(optimism)

$$\le J_{M_K}(\pi^*_{M_K}) - J_M(\pi^*_{\hat{M}_K})$$

$$\le J_{\hat{M}_K}(\pi^*_{\hat{M}_K}) + dis(M_K, \hat{M}_K)\cdot \frac{V_{max}}{2(1-\gamma)} \quad \text{(Lemma 2)} - J_M(\pi^*_{\hat{M}_K})$$

$$\le J_{M_K}(\pi^*_{\hat{M}_K}) + dis(M_K, \hat{M}_K)\cdot \frac{V_{max}}{1-\gamma} - J_M(\pi^*_{\hat{M}_K})$$

$$\le P_M[\text{escape} \mid \pi^*_{\hat{M}_K}]\cdot V_{max} + dis(M_K, \hat{M}_K)\cdot \frac{V_{max}}{1-\gamma}$$

$$\Downarrow$$

$$\varepsilon < P_M[\text{escape} \mid \pi^*_{\hat{M}_K}] + dis(M_K, \hat{M}_K)/(1-\gamma)$$

$\pi^*_{\hat{M}_K}$: exploration policy

$dis(M_K, \hat{M}_K)$: controlled by m.

then set $m = \tilde{O}\left(\frac{|S|}{\varepsilon^2(1-\gamma)^2}\log\frac{1}{\delta}\right)$.

$\Rightarrow dis(M_k, \hat{M}_k) \le \varepsilon(1-\gamma)/2$.

$\Rightarrow \varepsilon < P_M[\text{escape}|\cdots] + \varepsilon/2$.

$\Rightarrow P_M[\text{escape}|\cdots] > \frac{\varepsilon}{2}$.

$\triangle$

$$\frac{|S\times A|\cdot m}{\varepsilon/2} \Rightarrow \tilde{O}\left(\frac{|S|^2|A|}{\varepsilon^3(1-\gamma)^2}\log\frac{1}{\delta}\right).$$

Remark: Stopping criterion: optimistic value

$$J_M(\pi_M^*) - J_M(\pi_{\hat{M}_k}^*) \le J_{\hat{M}_k}(\pi_{\hat{M}_k}^*) - J_M(\pi_{\hat{M}_k}^*)$$

if this is $\le O(\varepsilon)$. can stop.

$J_{\hat{M}_k}(\pi_{\hat{M}_k}^*)$: known

$J_M(\pi_{\hat{M}_k}^*)$: estimated by MC.