Rmax $M = (S, A, P, R, \gamma, d_0, H)$. R_{max} $\frac{1}{1-\gamma}$.

$J(\pi^*) - J(\hat{\pi}) \leq \varepsilon \cdot V_{max}$

How many ep do we need?

Algorithm: $n(s,a)$, $n(s,a,s')$ } counters. $\hat{P}(s'|s,a) = \frac{n(s,a,s')}{n(s,a)}$.

threshold m. $\hat{M}_K = (S, A, \hat{P}_K, R_K, \gamma)$.

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

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

"Optimism in face of uncertainty".

Next ep: explore w/ $\pi^*_{\hat{M}_K}$.

in the next ep data. if any $(s_h, a_h) \notin K$.

$s_1, a_1, s_2, a_2, \ldots, s_H, a_H$

inc $n(s_h, a_h)$

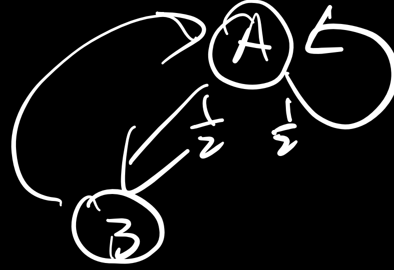
$n(s_h, a_h, s_{h+1})$.

"tape" / "cache".

T_A.

A B A B B - - -

fixed. → T_A (≠ / ∥)



A A A B A B A A - - -

r.v. → $\boxed{T}$ △.

~~$r + \gamma \max_{a'} f(s', a')$~~.

Table 1: Relationship between M, M_K, and $\widehat{M}_K$.

	M	M_K	$\widehat{M}_K$
Known (K)	$= M$	$= M$	$\approx M$
Unknown	$= M$	self-loop	self-loop

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

---

w.p. $\geq 1 - \delta$, $\forall (s,a) \in K$, $\| \hat{P}_K(\cdot | s,a) - P_K(\cdot | s,a) \|_1$

$\leq O\left(\sqrt{\frac{|S|}{m} \log \frac{|S \times A|}{\delta}}\right)$. (= ϵ_P.)

$\Rightarrow \forall \pi : S \to A$.

$\to \| V^{\pi}_{M_K} - V^{\pi}_{\hat{M}_K} \|_\infty \leq \epsilon_P \cdot \frac{V_{max}}{1 - \gamma}$

$$\rightarrow \| V^*_{M_K} - V^*_{\hat{M}_K} \|_\infty \leq \varepsilon_p \cdot \frac{V_{max}}{1-\gamma}.$$

"Explore-or-Terminate"

Fact: optimism:

$\forall \pi : S \to A.$

$J_M(\pi) \leq J_{M_K}(\pi)$

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

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

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

$$\leq J^*_{M_K} - J_M(\pi^*_{\hat{M}_K}) + \frac{\varepsilon_p V_{max}}{1-\gamma}$$

$$= J^{\hat{\pi}}_{M_K} - J_M(\hat{\pi}) + \frac{\varepsilon_p V_{max}}{1-\gamma}$$

$$\frac{\varepsilon V_{max}}{2} \leq J^{\hat{\pi}}_{M_K} - J_M(\hat{\pi}) + \frac{2\varepsilon_p V_{max}}{1-\gamma}$$

$$\frac{\varepsilon V_{max}}{2} < \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d^{\hat{\pi}}_{M_K}} \left[Q^{\hat{\pi}}_{M_K} - \mathcal{T}^{\hat{\pi}}_M Q^{\hat{\pi}}_{M_K} \right] + \frac{\varepsilon V_{max}}{2}$$

$$= \sum_{(s,a) \in K} d_M^{\hat{\pi}}(s,a) \cdot (0.) + \sum_{(s,a) \notin K} d_M^{\hat{\pi}}(s,a) \cdot (c V_{max})$$

$$Q_{M_K}^{\hat{\pi}} - T_M^{\hat{\pi}} Q_{M_K}^{\hat{\pi}} = T_{M_K}^{\hat{\pi}} Q_{M_K}^{\hat{\pi}} - T_M^{\hat{\pi}} Q_{M_K}^{\hat{\pi}}$$