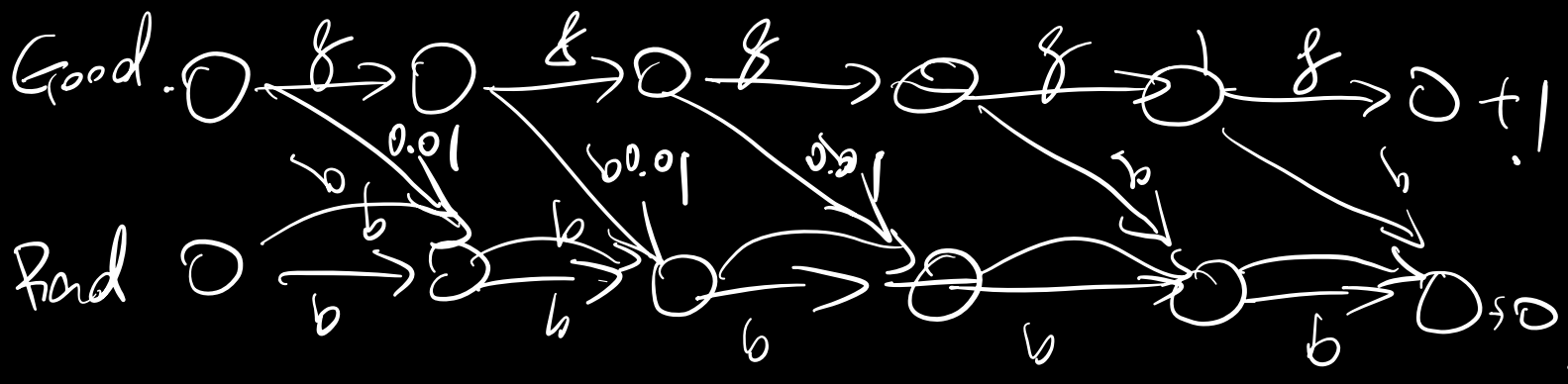
Rmax Exploration | "Combination Lock".



$$\nabla J(\pi) \propto \frac{1}{1-\gamma} \mathbb{E}_{d^\pi}\Big[\nabla \log \pi(a|s) \cdot Q^\pi(\quad) \sum_{s',r} r_{s'} \Big]$$

Want: sample complexity $\mathrm{poly}(|S|, |A|, H, \frac{1}{\varepsilon}, \frac{1}{\delta})$ ($H \approx \frac{1}{1-\gamma}$)

output $\hat{\pi}$ s.t. $J_M(\pi^*) - J_M(\hat{\pi}) \le \varepsilon \cdot V_{max}$ ||

Rmax. At any point of exploration.

Maintain.

- $n(s,a)$: # times see (s,a)
- $n(s,a,s')$: # times (s,a,s')

parameter.

Define $K := \{(s,a): n(s,a) = m\}$. ("known".)

Build MDP $\hat{M}_K$:

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

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

Collect next episode w/ $\pi^*_{\hat{M}_K}$.

Optimism in face of uncertainty

Analysis: Define M_K similar to $\hat{M}_K$.
except $P_K(s'|s,a) = P(s'|s,a) \quad \forall (s,a) \in K$.

	M	M_K	$\hat{M}_K$
Known(K)	$= M$	$= M$	$\approx M$
Unknown	$= M$	R_{max}-loop	R_{max}-loop

1. Optimism: $\forall \pi: S \to \Delta(A)$. $V^{\pi}_{M_K}(s) \geq V^{\pi}_{M}(s)$.

2. (Induced Ineq)
Given M_1, M_2, agree on $K \subseteq (S \times A)$.
Define $\text{escape}_K(\tau) = \begin{cases} 1, & \text{if } \tau \text{ visits } (s,a) \notin K \\ 0 \end{cases}$

$\forall \pi: S \to \Delta(A)$.

$$|J_{M_1}(\pi) - J_{M_2}(\pi)| \leq V_{max} \cdot P^{\pi}_{M_1}[\text{escape}(\tau)]$$

$\mathbb{P}_M^{\pi^*_{\hat{M}_K}}[\text{escape}(\tau)] \geq O(\varepsilon)$ → suboptimality.

---

mSA

ε

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

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

$$\leq J_{M_K}(\pi^*_{M_K}) - J_M(\pi^*_{\hat{M}_K}) \quad O\left(\sqrt{\tfrac{1}{m}}\right)$$

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

$$\leq V_{max} \cdot \mathbb{P}_M^{\pi^*_{\hat{M}_K}}[\text{escape}_K(\tau)] + \text{small}$$

$$\geq \frac{\varepsilon V_{max}}{2} \qquad \frac{\varepsilon V_{max}}{2}$$

$$\Rightarrow \mathbb{P}[\text{escape}] \geq \frac{\varepsilon}{2}.$$

Proof of induced ineq.

$$\left| J_{M_1}(\pi) - J_{M_2}(\pi) \right| \qquad f(s_0) - J_M(\pi) = \mathbb{E}_{d_M^\pi}[f - \mathcal{T}^\pi f]$$

$$= \left| J_{M_1}(\pi) - Q^\pi_{M_2}(s_0, \pi) \right|. \qquad (1-\gamma$$

$$= \frac{1}{1-\gamma} \left| \mathbb{E}_{d^\pi_{M_1}} \left[Q^\pi_{M_2} - \mathcal{T}^\pi_{M_1} Q^\pi_{M_2} \right] \right|.$$

$$\sum_{(s,a) \in K} d^\pi_{M_1}(s,a) \cdot \left(Q^\pi_{M_2} - \mathcal{T}^\pi_{M_2} Q^\pi_{M_2} \right)$$

$$\sum_{(s,a) \notin K} d^\pi_{M_1}(s,a) \cdot \left(\quad \right) \cdot$$

$O(V_{max})$