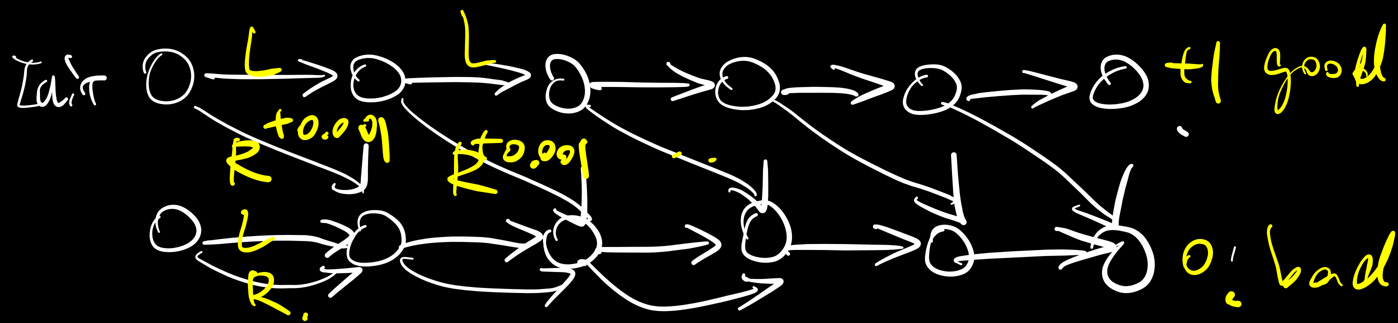
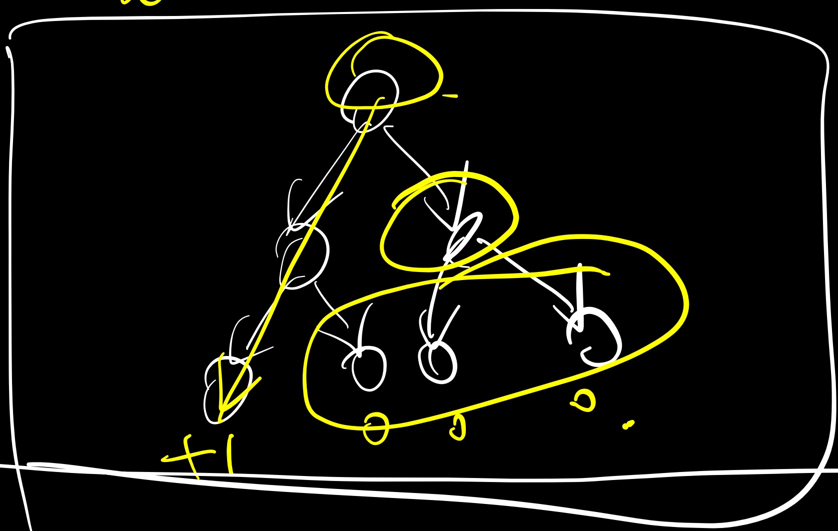


$$\mathbb{E}_{d^{\pi}}\left[\nabla \log \pi(s,a) \sum r_t \right].$$

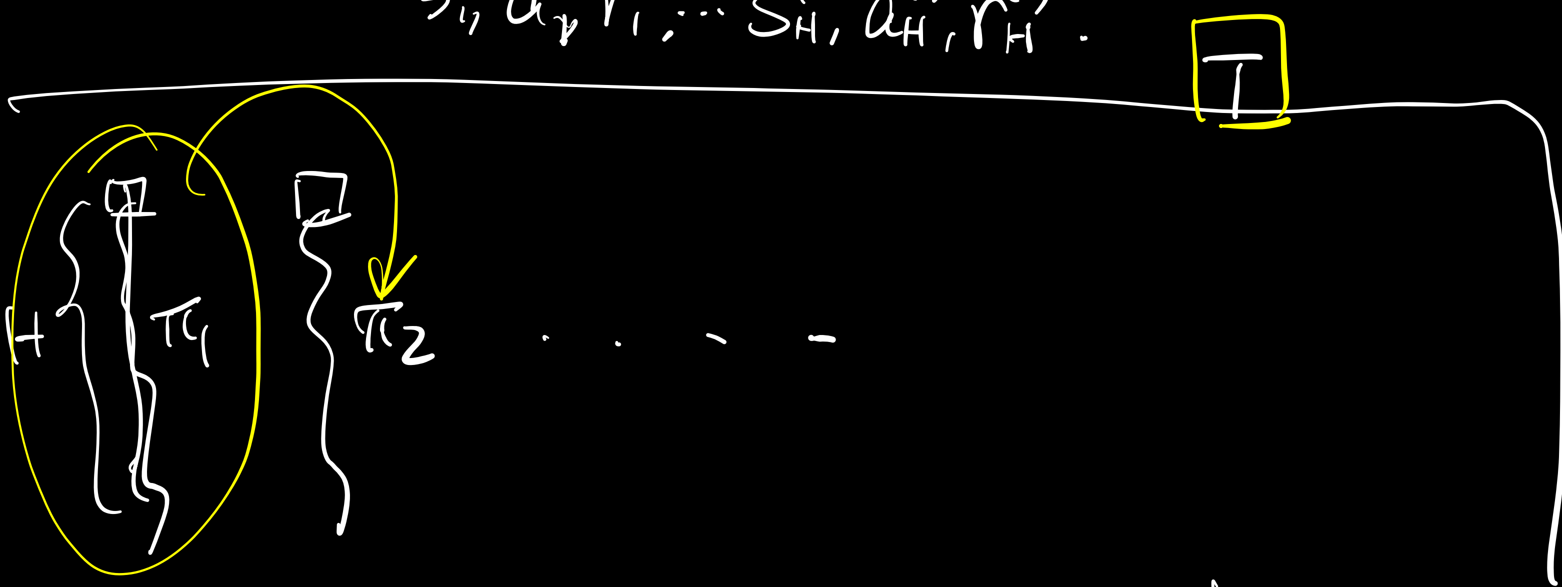


Online RL

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

- alg picks π_t.
- run π_t to collect traj.

$$s_1^{(t)}, a_1^{(t)}, r_1^{(t)}, \cdots, s_H^{(t)}, a_H^{(t)}, r_H^{(t)}.$$



Sample complexity | Alg: output $\hat{\pi}$. | $1/\varepsilon^2$.

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

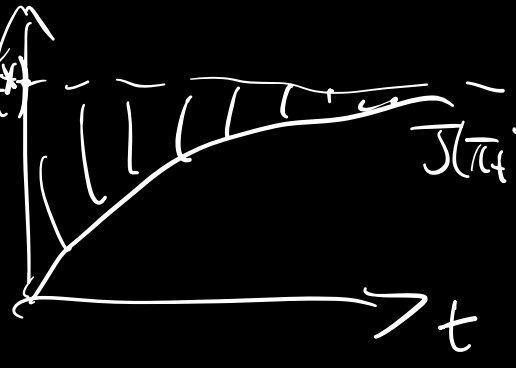
$$T = \text{poly}(|S|, |A|, H, 1/\varepsilon, \log 1/\delta).$$

Regret bound. $\text{Regret}_T = \sum_{t=1}^{T} J(\pi^*) - J(\pi_t)$.

Guarantee: w.p. $\geq 1-\delta$.

$$\text{Regret}_T = \text{poly}(S, A, H, \log 1/\delta)\, o(T).$$

$\sqrt{T}$. $T^{2/3}$ $T^{3/4}$.



Sample comp → regret.

Alg: $1/\varepsilon^2$.

$$\sum_{t=1}^{T} J(\pi^*) - J(\pi_t)$$

$$\leq T_0 \cdot V_{\max} + \underbrace{(T - T_0 + 1)}\left(J(\pi^*) - J(\hat{\pi})\right).$$

$$\leq T_0 V_{max} + \frac{T}{T_0} \cdot c \cdot \frac{1}{\sqrt{T_0}} \cdot$$

$$T_0 V_{max} + \frac{cT}{2\sqrt{T_0}} + \frac{cT}{2\sqrt{T_0}}$$

achieved by

$$T_0 V_{max} = \frac{cT}{2\sqrt{T_0}}$$

$$3\sqrt[3]{T_0 V_{max} \cdot \frac{cT}{2\sqrt{T_0}} \cdot \frac{cT}{2\sqrt{T_0}}} = 3\sqrt[3]{\frac{c^2 T^2 V_{max}}{4}}$$

$$2T_0^{3/2} = \frac{cT}{V_{max}}$$

$$T_0 = \left(\frac{cT}{2V_{max}}\right)^{2/3}.$$

---

regret $\rightarrow$ SC. run regret_min for T.

Output:

draw $i \sim Unif([T])$.

run π_i.

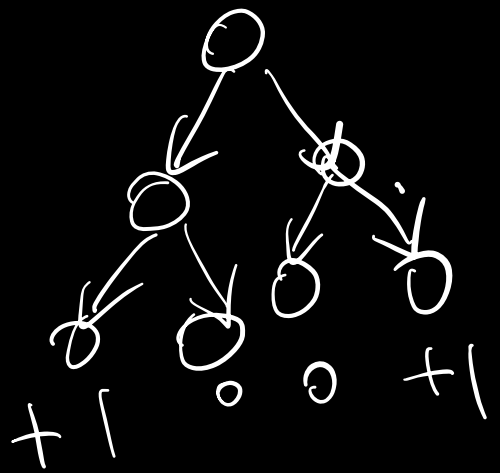
$[T] \rightarrow 1, 2, \cdots T.$

$$J(a^*) - \frac{\sum_{i=1}^{T} J(\pi_i)}{T} = \frac{1}{T} \sum_{i=1}^{T} J(a^*) - J(\pi_i)$$

$$= \frac{1}{T} O(\sqrt{T}).$$

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

$= O\left(\frac{1}{\sqrt{T}}\right)$.



L.

R.

$\pi_1(\cdot|s_1)$. $\pi_2(\cdot|s_1)$.

π_1 π_2 $1/2$, $1/2$.

$a_1 \sim \frac{\pi_1(\cdot|s_1) + \pi_2(\cdot|s_2)}{2}$.

s_1, a_1, s_2 $\boxed{a_2}$?

$\Pr\left(i \middle| s_1, a_1, s_2\right)$.