Online RL $M = (S, A, P, R, H, d_0)$.

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

- Learner chooses π_t to collect a traj.
- $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)$. "pure exploration"

After T rounds, learner output $\hat{\pi}$.

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

"simple regret". $= f(\frac{1}{T})$.

to achieve ε. How large $T = g(\frac{1}{\varepsilon})$ needs to be.

"sample complexity"

Regret: ("cumulative") regret $= V_{max} \cdot T$

$$\text{Regret} := \sum_{t=1}^{T} \left(J(\pi^*) - J(\pi_t) \right) = o(T).$$

"exploration - exploitation": $\sqrt{T}$.

tabular MDP: $\sqrt{HSAT}$

Sample comp $\Rightarrow$ regret.

$\hat{\pi}$ $\quad J(\pi^*) - J(\hat{\pi}) = O\left(\frac{1}{\sqrt{T}}\right)$.

$\pi_1, \pi_2, \cdots, \pi_T$.

(1) Run SC alg for $T_1 \le T$ episodes

(2) $\pi_{T_1:T} = \hat{\pi}$.

Regret $= T_1 + (T - T_1) \cdot \sqrt{\frac{1}{T_1}}$.

$\le T_1 + T\sqrt{\frac{1}{T_1}} = O(T^{2/3})$.

$x + T\sqrt{\frac{1}{x}} = x + \frac{1}{2}\sqrt{\frac{1}{x}} + \frac{1}{2}\sqrt{\frac{1}{x}}$

$\ge \sqrt[3]{x \cdot \frac{1}{2}\sqrt{\frac{1}{x}} \cdot \frac{1}{2}\sqrt{\frac{1}{x}}}$ $\quad T_1 = O(T^{2/3})$.

$= \sqrt[3]{\frac{1}{4}} = \Omega(T^{2/3})$.

Regret $\Rightarrow$ SC. $\pi_1, \pi_2, \cdots, \pi_T$

$$\sum_{t=1}^{T} \left(J(\pi^*) - J(\pi_t) \right) = O(\sqrt{T}).$$

$\hat{\pi}$ = unif mixture of $\pi_1, \pi_2, \cdots, \pi_T$.

i.e. draw $t \in \text{unif}\{1, \cdots, T\}$,
then run π_t.

$$J(\hat{\pi}) = \frac{1}{T} \sum_{t=1}^{T} J(\pi_t)$$

$$J(\pi^*) - J(\hat{\pi}) = \frac{1}{T} \left(\sum \left(J(\pi^*) - J(\pi_t) \right) \right)$$

$$= \frac{1}{T} \sqrt{T} = \frac{1}{\sqrt{T}}.$$