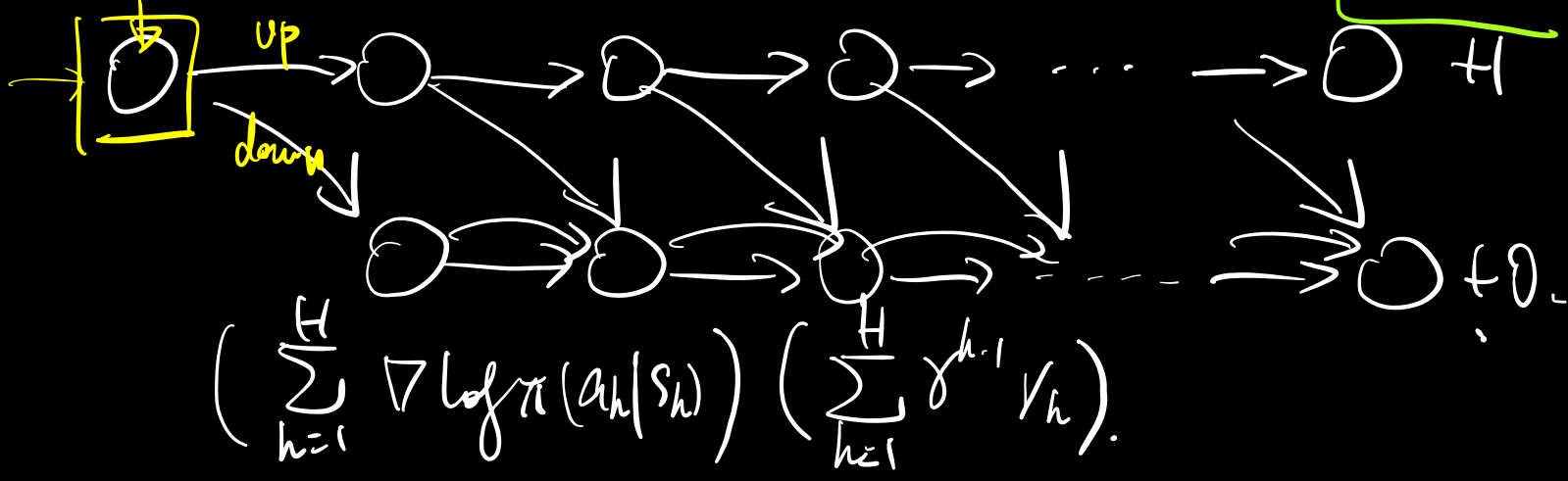
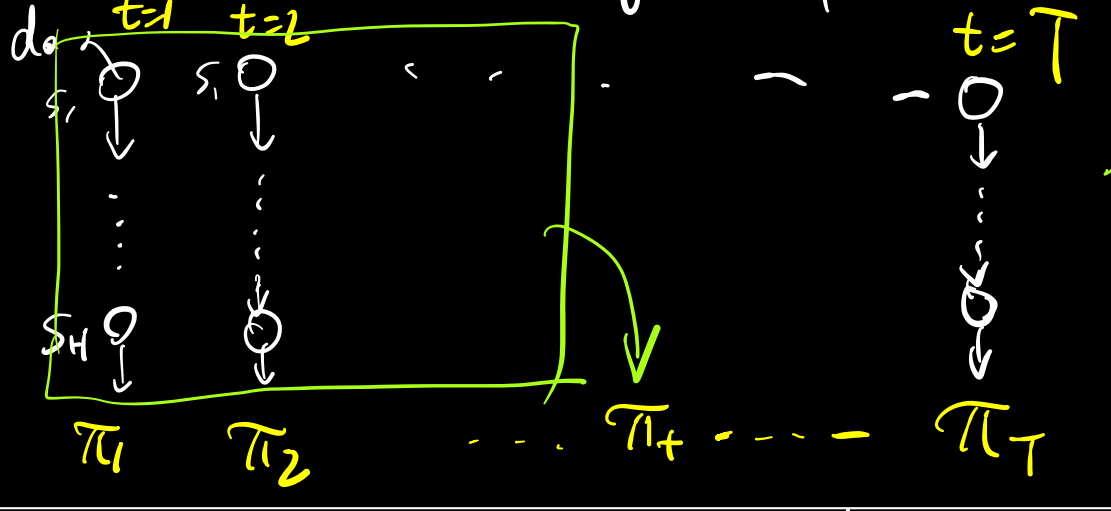
Exploration & Online RL

Combination lock. 2^{-H}



$$\left(\sum_{h=1}^{H} \nabla \log \pi(a_h|s_h)\right)\left(\sum_{h=1}^{H} \gamma^{h-1} r_h\right).$$

Finite-horizon setting: for $t = 1, 2, 3, \ldots, T$.



$\hat{\pi}$

$J(\pi^*) - J(\hat{\pi}) \le \epsilon$.

Sample complexity (pure exploration): w.p. $\ge 1-\delta$.

$J(\pi^*) - J(\hat{\pi}) \le \epsilon$, as long as $T =$ some fn of $1/\epsilon$ (and δ, $|S|$, $|A|$, H, ...).

$O(1/\epsilon^2)$

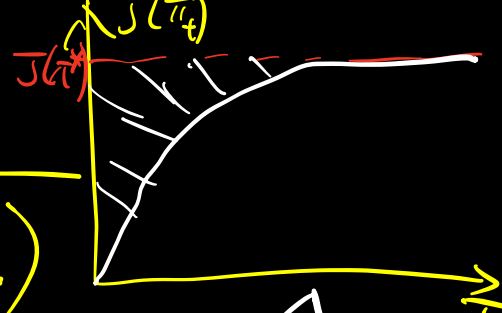
Regret (exploration & exploitation).

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

w.p. $\ge 1-\delta$, $\text{Regret}_T = o(T)$.

$\Rightarrow \text{Regret}_T / T = o(1)$.

$O(\sqrt{T})$



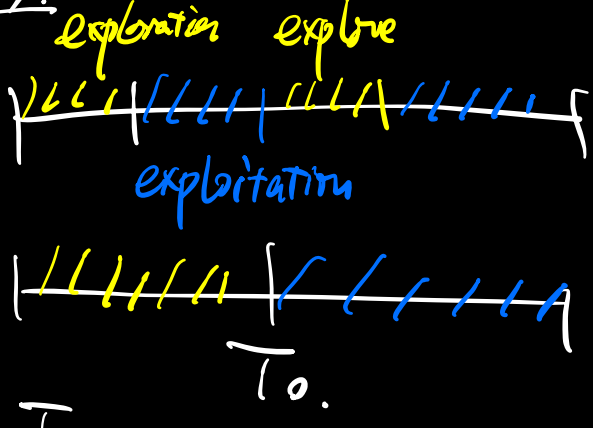
Pure-explore $\Rightarrow$ regret-min (exp-exp).

Alg A. s.t. $T = O(1/\varepsilon^2)$.

Reduction: ① run A for T_0 steps

② get $\hat{\pi}$.

③ deploy $\hat{\pi}$ for $T - T_0$.



$$\text{Regret}_T = \sum_{t=1}^{T_0} \left(J(\pi^*) - J(\pi_t) \right) + \sum_{t=T_0+1}^{T} \left(J(\pi^*) - J(\hat{\pi}) \right)$$

$$\leq V_{max} \cdot T_0 + T \times C \cdot \sqrt{\frac{1}{T_0}}.$$

$$V_{max} \cdot T_0 + \frac{CT}{2}\sqrt{\frac{1}{T_0}} + \frac{CT}{2}\sqrt{\frac{1}{T_0}} \geq 3\sqrt[3]{V_{max} \cdot C^2 \cdot T^2}$$

$$V_{max} T_0 = \frac{CT}{2}\sqrt{\frac{1}{T_0}} \Rightarrow = O(T^{2/3}).$$

Regret-min $\Rightarrow$ sample complexity

Alg A: $\pi_{1:T}$. $\sum_{t=1}^{T} \left(J(\pi^*) - J(\pi_t) \right) = c\sqrt{T}$.

want: $\hat{\pi}$ $J(\pi^*) - J(\hat{\pi}) \leq \varepsilon$.

$\hat{\pi} = \mathrm{Unif}(\pi_{1:T})$: (traj-level)

① At $h=1$, Choose $i \sim \mathrm{unif}([T])$

② exec. π_i for H steps

$\frac{\pi_1+\pi_2}{2}$: $\pi(a|s) = \frac{\pi_1(a|s)+\pi_2(a|s)}{2}$

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

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

$$= \frac{c\sqrt{T}}{T}$$