Exploration

RL settings according to how to interact w/ env.

→ batch : no interactions. given data (collected by someone else).

→ planning: have access to a "generative model".
i.e. can query $s' \sim P(\cdot \mid s,a)\ \forall s,a$.
"sample complexity" → "query complexity".

MCTS.
tabular.

→ online : can interact w/ env. in terms of trajectories.
weaker than "gen model".

---

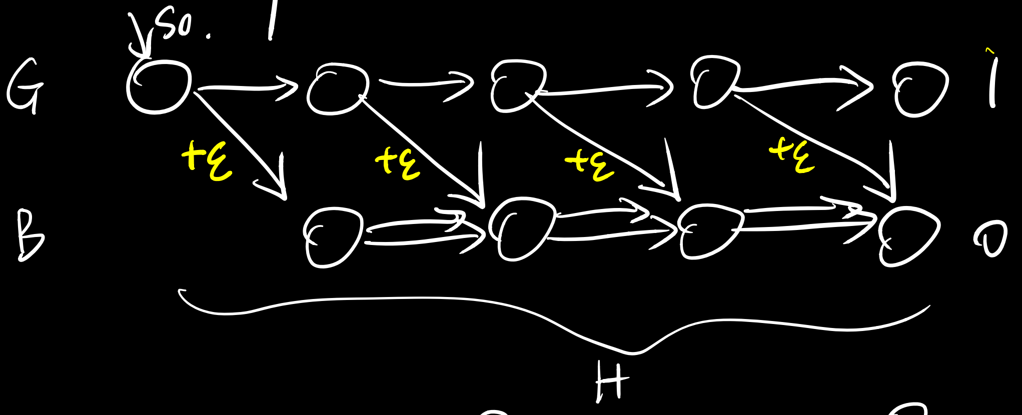
online setting: learner can interact w/ env.
w.t. learn ε-optimal policy : how many trajectories need to be drawn?

Typical exploration strategy:
1. learning component
2. exploration component

DQN + ε-greedy exploration.
"learning component"
$\begin{cases} \text{w.p. } 1-\varepsilon, \ \pi_{\hat{Q}} \\ \text{w.p. } \varepsilon. \ \ \text{unif}(A). \end{cases}$

Unif exploration has exponential sample complexity:

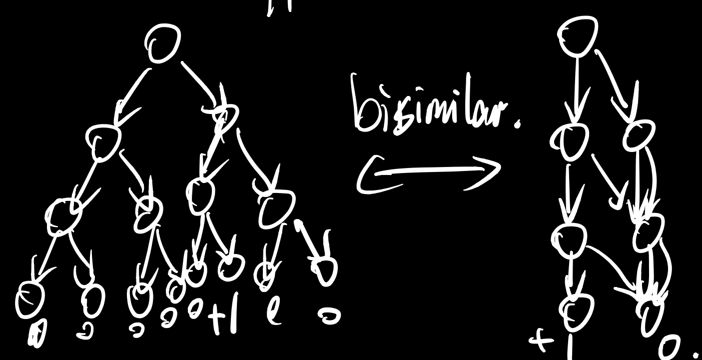
Counterexample: "Combination lock".



$\Omega(1/2^H)$ prob.

+greedy doesn't.
① if $\hat{Q}$ is init as 0, then no update.
② "anti-shaping".

Fun fact:



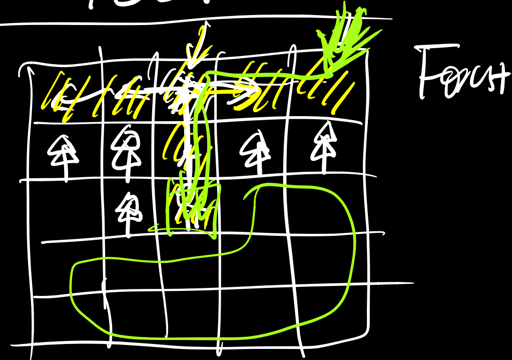
Q: Can we provably explore w/ poly sample complexity, at least in tabular RL? A: YES!

Formal setup: $M = (S, A, P, R, \gamma, d_0)$.

episodic: all traj. terminates in H steps.

find $\hat{\pi}$, s.t. $J(\pi^*) - J(\hat{\pi}) \le \varepsilon \cdot V_{max}$.

$J(\pi) := \mathbb{E}_{s \sim d_0}[V^\pi(s)]$.



$V_{max} := \dfrac{R_{max}}{1-\gamma}$.

Rmax-alg: alg maintains: $\forall s, a, s'$, $\begin{cases} n(s,a) & \text{(initially: 0)}. \\ n(s,a,s'). \end{cases}$

hyperparam: m. ("threshold")

$K := \{(s,a) \in S \times A : \underline{n(s,a) = m}\}$. "optimism in face of uncertainty" (OFU)

1. Build MDP $\hat{M}_K$: $\forall s, a, s'$

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

$$\hat{R}_K(s,a) = \begin{cases} R(s,a), & \text{if } (s,a) \in K. \\ R_{max}. & \text{o.w.} \end{cases}$$

Stop criterion?

2. Collect a traj: $s_1, a_1, r_1, \dots, s_H, a_H, r_H$ using $\pi^*_{\hat{M}_K}$

3. $\forall h$ s.t. $n(s_h, a_h) < m$, $\Rightarrow$ $n(s_h,a_h)$++, $n(s_h, a_h, s_{h+1})$++.

Analysis. $d^\pi_M(s,a) := (1-\gamma) \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{P}_M[s_h = s, a_h = a \mid \pi]$.

Define M_K as "expected ver" of $\hat{M}_K$.

$$P_K(s'|s,a) = \begin{cases} P(s'|s,a), & \text{if } (s,a) \in K \\ \mathbb{I}(s'=s) & \text{o.w.} \end{cases} \qquad R_K(s,a) = \begin{cases} R(s,a), & \text{if } (s,a) \in K \\ R_{max}. & \text{o.w.} \end{cases}$$

Def: $\Delta(M_1, M_2) := \max_{s,a} \|P_1(s,a) - P_2(s,a)\|_1$.

	M	M_K	$\hat{M}_K$
K	$=M$	$=M$	$\simeq M$
unknown	$=M$	loop	loop

Claim: $\Delta(M_K, \hat{M}_K)$ is "small".
(as a function of m).

$M_K \approx \hat{M}_K$

Δ

Lemma: fix (s,a) where $n(s,a) = m$, w.p $\geq 1-\delta$.

$$\|\hat{P}_K(\cdot|s,a) - P_K(\cdot|s,a)\|_1 \leq \sqrt{\frac{2}{m}\log\frac{2(2^{|S|})}{\delta}}.$$

Coro: w.p $\geq 1-\delta$, $\Delta(M_K, \hat{M}_K) \leq \sqrt{\frac{2}{m}\log\frac{2\cdot 2^{|S|}\cdot |S||A|}{\delta}}$

Lemma (optimism): $\forall \pi: S \to A$, $J_{M_K}(\pi) \geq J_M(\pi)$.

Further implications of $\hat{M}_K \approx M_K$.

"Simulation Lemma"

① $\forall \pi$, $|J_{\hat{M}_K}(\pi) - J_{M_K}(\pi)| \leq \Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{2(1-\gamma)}$.

② $\|V^*_{M_K} - V^*_{\hat{M}_K}\|_\infty \leq \Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{2(1-\gamma)}$.

Let T_K and $\hat{T}_K$ be the Bellman op in M_K, $\hat{M}_K$ resp.

$$\|V^*_{M_K} - V^*_{\hat{M}_K}\|_\infty = \|V^*_{M_K} - \hat{T}_K V^*_{M_K} + \hat{T}_K V^*_{M_K} - \hat{T}_K V^*_{\hat{M}_K}\|_\infty$$

$$\leq \|V^*_{M_K} - \hat{T}_K V^*_{M_K}\|_\infty + \gamma \|V^*_{M_K} - V^*_{\hat{M}_K}\|_\infty.$$

$$\|V^*_{M_K} - \hat{T}_K V^*_{M_K}\|_\infty = \|T_K V^*_{M_K} - \hat{T}_K V^*_{M_K}\|_\infty.$$

$$= \gamma \max_{s,a} \left| \mathbb{E}_{s' \sim P_K(\cdot|s,a)}[V^*_{M_K}(s')] - \mathbb{E}_{s' \sim \hat{P}_K(\cdot|s,a)}[V^*_{M_K}(s')] \right|$$

$$= \gamma \max_{s,a} \left| \langle P_K(s,a) - \hat{P}_K(s,a), V^*_{M_K} - \frac{V_{max}}{2}\cdot \mathbf{1} \rangle \right|$$

$$\le \gamma \max_{s,a} \Big(\underbrace{\| P_K(s,a) - \hat{P}_K(s,a) \|_1}_{\mathbb{R}^S} \cdot \frac{V_{max}}{2} \Big).$$

$$= \gamma \cdot \Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{2}.$$

$\| T_1 - T_2 \| = \max_V | T_1 V - T_2 V |$

$\boxed{\text{Key Lemma: "Induced Ineq"}}$ $\forall \underline{\pi}: S \to A$.

- $|J_M(\pi) - J_{M_K}(\pi)| \le V_{max} \cdot \underbrace{\mathbb{P}_M[\text{escape}_K(\tau) | \pi]}_{\Delta}$.
- $\mathbb{P}_M[\cdot_\Delta | \pi]$ considers dist. of rand traj. generated in M.
- τ is rand. traj. $\tau = (s_1, a_1, s_2, a_2, \dots, s_H, a_H)$.
- $\text{escape}_K(\cdot) = 1$ if $\exists h$. $s_h, a_h \notin K$.

$S \times A$, K, $\cup K$

Proof: $\underline{J_{M_K}(\pi)} = \sum_{\tau: \text{escape}_K(\tau)=1} \underline{\mathbb{P}_{M_K}[\tau|\pi] R(\tau)}$.

$$+ \sum_{\tau: \text{escape}_K(\tau)=0} \mathbb{P}_{M_K}[\tau|\pi] R(\tau).$$

$\to R(\tau) = \sum_{h=1}^{H} \gamma^{h-1} R(s_h, a_h).$

$$= \sum_{\text{escape}_K(\tau)=1} \mathbb{P}_{M_K}[\tau|\pi] \cdot \big(\underline{R(\text{pre}_K(\tau))} + \underbrace{R(\text{suf}_K(\tau))}_{\Delta} \big)$$

$$+ \sum_{\text{escape}(\tau)=0} \mathbb{P}_{M_K}[\tau|\pi] R(\tau).$$

$\text{pre}_K \{ s_1, a_1, s_2, a_2$; $\text{suf}_K \{ s_3, a_3$

$$\le \sum_{\text{escape}_K(\tau)=1} \mathbb{P}_{M_K}(\tau|\pi) \cdot \underline{(R(\text{pre}_K(\tau)) + V_{max})}.$$

$$+ \cancel{\sum_{\text{escape}(\tau)=0} \mathbb{P}_{M_K}[\tau|\pi]} R(\tau).$$

$$= \sum_{\text{pre}_K(\tau)} \underbrace{\mathbb{P}_{M_K}[\text{pre}_K(\tau) | \pi]}_{} \big(\cancel{R(\text{pre}_K(\tau))} + V_{max} \big) +$$

Δ

$$J_{\underline{M}}(\pi) \geq \cancel{\sum_{pre_K(\tau)} \mathbb{P}_M[pre_K(\tau)|\pi]\, R(pre_K(\tau))}$$

$$+ \cancel{\sum_{escape_K(\tau)=0} \mathbb{P}_M[\tau|\pi] \cdot R(\tau)}.$$

$pre_K(\tau) = s_1, a_1, \cdots, s_h, a_h.$

s.t. $\forall t \leq h-1$, $(s_t, a_t) \in K$.

& $(s_h, a_h) \notin K$.

$\mathbb{P}_{M_K}(pre_K(\tau)|\pi).$

$= d_0(s_1) \cdot \pi(a_1|s_1) \cdot$

$P_K(s_2|s_1, a_1) \cdot \pi(a_2|s_2) \cdots$

$\cdots P_K(s_h|s_{h-1}, a_{h-1}) \cdot \pi(a_h|s_h)$

$P(s_2|s_1, a_1)$

Claim: $\mathbb{P}_{M_K}[pre_K(\tau)|\pi] = \mathbb{P}_M[pre_K(\tau)|\pi]$ Δ

$\forall\, escape_K(\tau)=0,\ \mathbb{P}_{M_K}[\tau|\pi] = \mathbb{P}_M[\tau|\pi]$.

$$J_{M_K}(\pi) - J_M(\tau) \leq \sum_{\substack{pre_K(\tau)\\ \Delta}} \mathbb{P}_M[pre_K(\tau)|\pi] \cdot V_{max}.$$

$$= \mathbb{P}_M[escape(\tau)|\pi]$$

Sample complexity analysis. $\forall$ episode. either of following happens.

① . $\pi^*_{\hat{M}_K}$ is ε-optimal $\leftarrow$ ✓ "terminate-or-explore".

② $\pi^*_{\hat{M}_K}$ escapes w/ significant prob. Δ

Proof: W.T.S. $\neg$ ① $\Rightarrow$ ②.

optimism

$$\boxed{\varepsilon \cdot V_{max}} < J_M(\pi^*) - J_M(\pi^*_{\hat{M}_K}) \leq J_{M_K}(\pi^*) - J_M(\pi^*_{\hat{M}_K}).$$

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

$$\leq J^*_{\hat{M}_K} + \Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{2(1-\gamma)} - J_M(\pi^*_{\hat{M}_K})$$

$= J_{\hat{M}_K}(\pi^*_{\hat{M}_K})$

$$\leq J_{M_K}(\pi^*_{\hat{M}_K}) + \Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{1-\gamma} - J_M(\pi^*_{\hat{M}_K}).$$

$$\le \underbrace{\Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{1-\gamma}}_{\Delta} + \underbrace{V_{max} \cdot \mathbb{P}_M[\text{escape}_K(\tau) \mid \pi^*_{\hat{M}_K}]}_{\Delta}.$$

$$\max_{s,a \in K} \| P_K(\cdot|s,a) - \hat{P}_K(\cdot|s,a) \|_1 \le \underbrace{\sqrt{\frac{2}{m} \log(\cdots)}}_{\Delta}.$$

Outline of remainder of proof:

Set m large enough, s.t. $\Delta(M_K, \hat{M}_K) \cdot \frac{V_{max}}{1-\gamma} \le \frac{\varepsilon}{2} \cdot V_{max}$

$$\Rightarrow \mathbb{P}_M[\text{escape} \mid \pi^*_{\hat{M}_K}] \ge \frac{\varepsilon}{2}$$

$$m = \underline{\tilde{O}}\left(\frac{|S|}{\varepsilon^2 (1-\gamma)^2} \cdot \log\frac{1}{\delta}\right),$$

crude calc: $\underbrace{m \cdot |S| \times |A|}\ / \ \underbrace{\frac{\varepsilon}{2}}$

Remarks: ① can terminate if $\underline{J_M(\hat{\pi}^*_{M_K})} \approx \underline{J^*_{\hat{M}_K}}$.

can estimate from Monte-Carlo

② Rmax doesn't necessarily explore. ⇐

→ "reward-free exploration"

③ "hard-to-reach" states don't bother us.