Exploration and Learning in "Tabular" MDPs:

$\Delta(\mathcal{X}) = \{$probability dist. on $\mathcal{X}\}$

What is an MDP? (finite horizon H, episodic)

(S, A, P_h, R_h, d_0) (for $h = 1 \ldots$ to H)

$P_h : S \times A \to \Delta(S)$

$R_h : S \times A \to \Delta(\mathbb{R})$

Policy: $\pi_h : S \longrightarrow \Delta(A)$ for $h=1$ to H.

π^*: optimal policy

Tabular: $|S|$ and $|A|$ are not too big

Given MDP, find π^* via Bellman Equations.

$$Q_h^*(s,a) = \mathbb{E}_{\substack{s_{h+1} \sim P_h(s,a) \\ r_h \sim R_h(s)}} \left[r_h + V_{h+1}^*(s_{h+1}) \right]$$

$$V_h^*(s) = \max_a Q_h^*(s,a)$$

$$\pi_h^*(s) = \arg\max_a Q_h^*(s,a)$$

Problem for Reinforcement Learning:

oftentimes P_h, R_h are unknown.

BUT: given Q_h^*, can compute π_h^*.

Q: How do we find Q_h^*? (or π_h^*?)

Today: Tabular MDP where P_h, R_h etc unknown (e.g. Q_h^*)

R_h: reward function is unknown,

but can usually deal with it using UCB.

(like bandits)

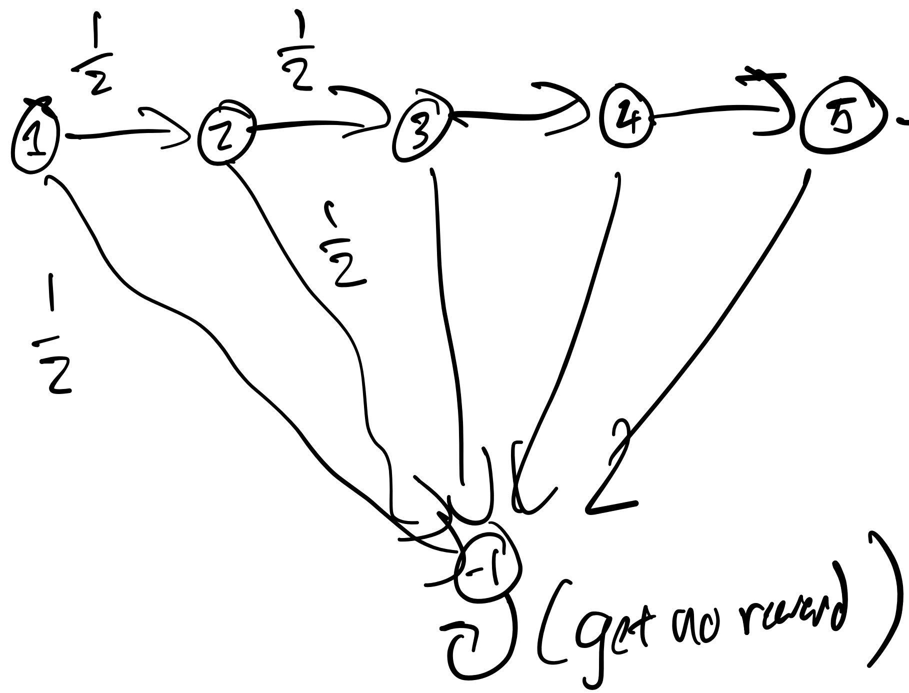
So assume R_h is known. (and d_0)

For Tabular MDP, main issue is P_h is unknown.

Want to learn P_h by playing many episodes.

Q: Can we learn $P_h(s,a)$ for all $(s,a) \in S \times A$?

P: Some states may be hard to reach.



$\mathcal{A} = \{\text{left}, \text{right}\}$

$\pi^*_h(s)$ unknown
randomly
fixed by nature

$R^*_h(s,a) = \begin{cases} 1 & \text{if } a = \pi^*(s) \\ 0 & \text{o.w.} \end{cases}$

for every $s \geq 1$, you pull either left or right arm

and if you pull "correct one"

For any policy

$$\Pr(s_H = H) = \left(\frac{1}{2}\right)^{H-1}.$$

RMAX algorithm:

"Optimism under uncertainty" $r_h \in [0,1]$ almost surely

- Maintain $K \subseteq S \times A$ parameters $\gg$ #states $= |S|$

$= \{(s,a) : n(s,a) \geq N\}$

- If $(s,a) \notin K$, we assume it is "maximally good" (optimism)

i.e. pretend $Q^*_h(s,a)$ "=" $H-h+1$.

Rmax algorithm: For episode $t=1$ to T:

- Compute Q^*_t under the (optimism) assumption for $(s,a) \notin K$.

For $(s,a) \in K$: Estimate $\hat{P}_h(s,a)$ from past experience.

- Play optimal strategy according to Q^*_t.
- $n(s,a) = n(s,a)+1$ for any (s,a) tried in this episode.

$$Q^*_{t,h}(s,a) = \mathbb{E}\left[r_h + V^*_{t,h+1}(s_{h+1}) \right]$$

$r_h \sim R_h(s,a)$

$s_h \sim \hat{P}_h(s,a)$

$\hat{P}_h(s_h, a_h) =$ Observed distribution over S from past.

FOR $(s,a) \in K$.

$V^*_{t,h}(s) = \max_a Q^*_{t,h}(s,a)$

(otherwise, $Q^*_{t,h}$ is optimistic).

Play $\pi_t^* \Longrightarrow$ observe $(s_1, a_1), (s_2, a_2), \ldots (s_H, a_H)$

Then $n(s_i, a_i) = n(s_i, a_i) + 1$ for all $i \in [H]$

$m(s_i, a_i, s_{i+1}) = m(s_i, a_i, s_{i+1}) + 1$

$n(s,a) = 0 \quad \forall S \times A$ initially.

$m(s, a, s') = 0$ initially

Really $K \subseteq \{(s,a,h) : n(s,a,h) \geq N\}$

$n(s_h, a_h, h) = n(s_h, a_h, h) + 1$

$m(s_h, a_h, h, s_{h+1}) = m(s_h, a_h, h, s_{h+1}) + 1.$

$$\hat{P}_h(s,a)(s_{h+1}) = \frac{m(s,a,h,s_{h+1})}{\sum_{s'} m(s,a,h,s')}.$$

<u>Thm</u>: With probability at least 99%, after $T = \mathrm{poly}(|S|, |A|, H, 1/\varepsilon)$ plays π^*_T is ε-optimal.

$(f(\pi^*_T) \geq f(\pi^*) - \varepsilon).$

Pf sketch of Thm:

At each time t, either

(A) π^*_t has $\gg \frac{\varepsilon}{H}$ probability of escaping $\mathcal{K}$

or

(B) π^*_t is ε-optimal.

(B) $\rightsquigarrow$ Done. (by optimism + $\mathcal{K}$ being well-observed)

(A) $\rightsquigarrow$ after $\sim \frac{H}{\varepsilon}$ rounds, some $n(s,a,h)$ some increases by 1 for $(s,a,h) \notin \mathcal{K}$.

SO $n(s,a,h) < N \Longleftrightarrow (s,a,h) \in \mathcal{K}$

$N \simeq \frac{H^2 |S| \log |S||A|H}{\varepsilon^2}$

case (A) can only occur roughly

$\approx \frac{H}{\varepsilon} |S||A|H$ many times

$d_{TV}(\hat{P}_h(s,a), P_h(s,a)) \leq \frac{\varepsilon}{H}$

(after this many times, $n(s,a,h) \geq N$ for all (s,a,h), i.e. $\mathcal{K} = \phi$)

Rigorous pf idea: (1) $Q^*_t = Q^{*, M_t}$ is true Q^* function for "optimistic MDP" M_t.

(2) If not in case (A), then w.p. $\approx 1-\varepsilon$ optimistic MDP and true MDP behave same.

Example of solvable non-Tabular MDPs?

action u_h insted of a_h

$$S = \mathbb{R}^d$$

Linear dynamics.

$$s_{h+1} = A s_h + B u_h + \text{noise}$$

u_h = "control input" given by policy

<u>Objective</u>:

$$R_h(s_h) = \frac{1}{2} \langle s_h, M_h s_h \rangle$$

objective

<u>Fact</u>: $u_h = K_h s_h$. LQR.