C 221 Lecture 8

Modeling

MDP

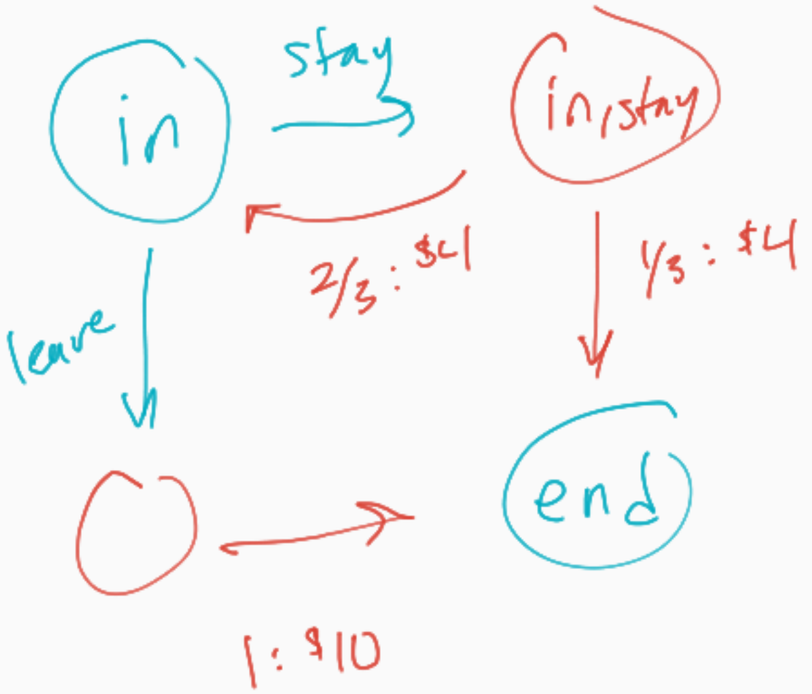
Inference

policy evaluation

value iteration

Learning

→ RL



$\pi(s) \rightarrow a$

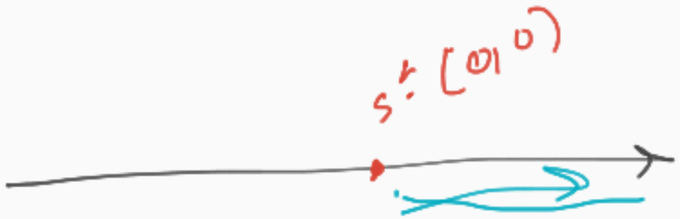
random path $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$

"episode" "roll-out"

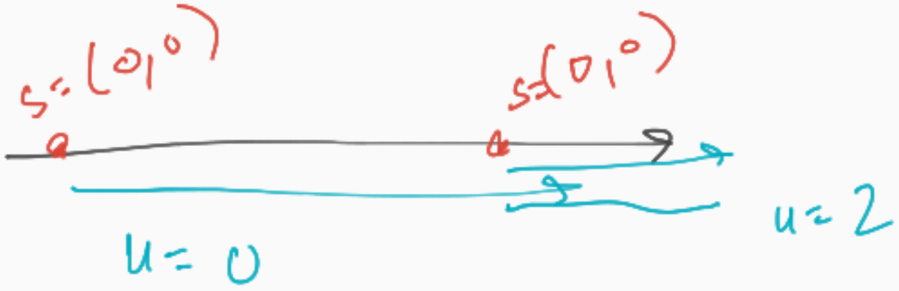
utility $u = r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots$

"return" / "discounted return"

$$\hat{Q}^{S}((0,0), a=0) = \frac{5+2+0}{3}$$



$$\gamma = 0.9$$



$$u_1 = 4 + 0 + 0$$

$$\hat{Q}_{TI} = \frac{u_1 + u_2 + u_3}{3}$$

$$u_2 = 4 + \gamma 4$$

$$u_3 = 4 + \gamma 4 + \gamma^2 4 + \gamma^3 4$$

Model-based value iteration $\hat{T}$ $\hat{R}$ eward $\rightarrow \hat{Q}^*$]

Model-free Monte Carlo $\hat{Q}_\pi$ $u \rightarrow \hat{Q}_\pi$

batch average, online convex stochastic
combination) gradient.

SARSA $\hat{Q}_\pi$ $r + \gamma \hat{Q}_\pi \rightarrow \hat{Q}_\pi$]

Q-learning $\hat{Q}^*$ $r + \gamma \hat{Q}^* \rightarrow \hat{Q}^*$]

$\uparrow$ max a'

exploration → epsilon-greedy

$u_1 = 4 \quad u_2 = 1 \quad u_3 = 1$

$\frac{4 + |-1|}{3} = 2$

$0 \cdot 0 + 1 \cdot 4$ $\quad \hat{Q}_\pi = 4$

$\frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 1$ $\quad \hat{Q} = 2 + 1/2$

$\frac{2}{3}(5/2) + \frac{1}{3} \cdot 1 = \frac{5}{3} + \frac{1}{3} = \frac{6}{3} = 2$