Reinforcement Learning

Final exam : Friday May 15th

- 7:15 PM - 10:15 PM
- Online
- Multiple choice
- Calculator needed
- Practice Gradiance Quiz will be posted
- One more Gradiance Quiz left (RL)

$\mathcal{A}$ - set of all actions

$A_t \in \mathcal{A}(S_t) \subset \mathcal{A}$

$R_t \in \mathbb{R}$

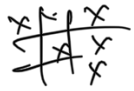
Policy π

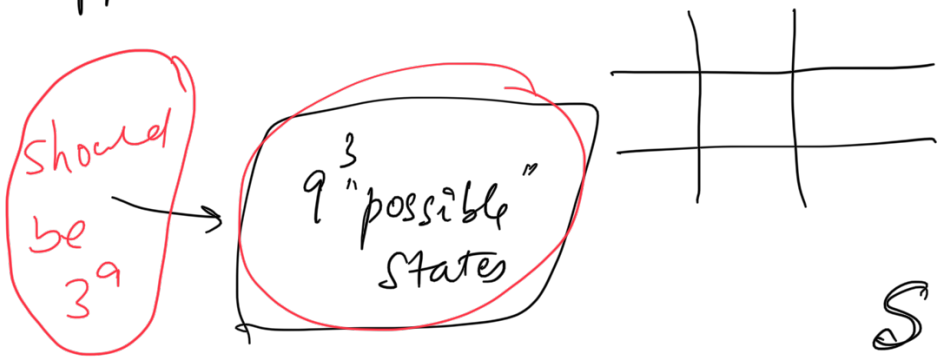
$S_1 \rightarrow A_{S_1}$

$S_2 \rightarrow S_2$

TicTacToe

Playing against an inperfect opponent.

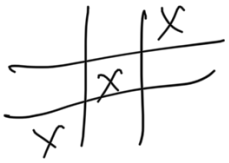


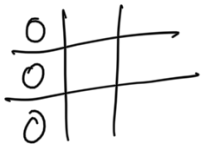


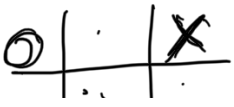
Value Prob of winning from that state.

S is small for ia.


0.5


1


0



$A(S_0)$

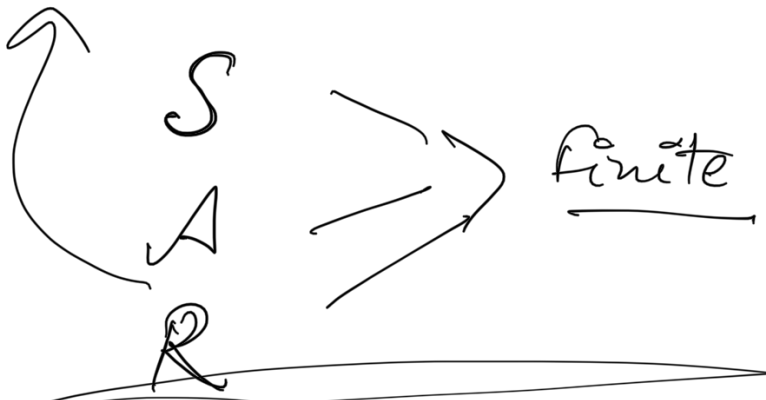
S_1

$0.5 + \alpha(0 - 0.5)$

$0.5 - 0.5\alpha$

Markov Decision processes (MDP)

Finite MDP



$$p(s', r \mid s, a) \doteq P(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a)$$

$|S| = 2 =$ on, off

$|A| = 4 = u, d, l, r$

$|R| = 2 = +1, -1$

$p(on, +1 \mid on, u)$

→ $p(s', r \mid s, a)$

$$p(s' \mid s, a) \doteq \sum_{r \in R} p(s', r \mid s, a)$$

G_t = Expected return from $t \rightarrow T$

Reinforcement learning - part II

Announcements

1. Gradiance is due on Sunday night
2. For finals, a practice quiz and list of topics will be posted soon
3. Assignment 3 due this week on Friday

RL

↳ Markov Decision Processes (MDP)

$$p(s', r \mid s, a)$$

$$G_t = R_{t+1} + \cdots$$

Agent

Policy (π)

$$\pi(a|s) \rightarrow A_t = a \mid S_t = s$$

value function at state s

$$V_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = s]$$

Complete model of the environment.

$$p(s', r \mid s, a)$$

Dynamic Programming (DP) based methods

Init: Start with a policy

Step 1. Find $V_\pi(s)$ $\forall s \in S$

Policy prediction

Step 2 $\pi \Longrightarrow V_\pi(s) \longrightarrow \pi'$

Policy Improvement

$p(s', r \mid s, a)$

DP → works if we ~~have a perfect~~ know the environment model completely $p(s', r | s, a)$

→ Data-free

→ not very useful

Monte-carlo methods.

Policy prediction

Policy Control

---

first visit MC prediction

S_1 $V(S_1) = 1$ $Returns(S_1) = [\,]$

S_2 $V(S_2) = 1$ $Returns(S_2) = [\,]$

$S_1 A_0 R_1 S_1 A_1 R_2 S_2 A_1 R_2 S_1 A \ldots$

$V(S_t)$

$Q(S_t, a_t)$

PlayerTotal, AceUsable, HouseCard

S_1	11	0	6
S_2	13	0	6
S_3	17	0	6
S_4	18	0	6
S_5	22	0	6

Returns $(22, 0, 6) = [-1]$

Returns $(18, 0, 6) = [\ -\ ]$

$V(\_\_) = -1$,

+1	0	6