687 - Lecture 2. Sept 7^{th} 2017

MDP

- A mathematical formulation of the environment and what we want the agent to learn.

- $t \in \{0, 1, \ldots\}$ time step

- S_t, A_t, R_t ← Reward given to the agent at time t.

S_t ↑ State at time t

A_t ↑ Action chosen at time t

MDP $M = (\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$

1. $\mathcal{S}$ = Set of all possible states of the env. [Finite]
2. $\mathcal{A}$ = Set of all possible actions [Finite]
3. P = "Transition function" describes how states of the env. transition.

Finite assumption for MDP (as opposed to continuous)

$$P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \Rightarrow [0, 1]$$

$$p(s, a, s') \triangleq \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$$

↑ [Is defined to be]

$\forall s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$

4. R: "Reward function" describes how rewards are generated. $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$.

$$R(s, a, s') = \mathbf{E}[R_t \mid S_t = s, A_t = a, S_{t+1} = s']$$

$p(8, \text{right}, 9) = 0.8$

$p(8, \text{right}, 15) = 0.$

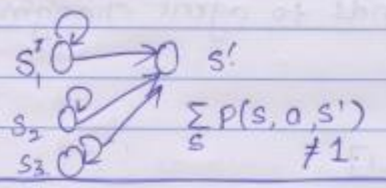
Deterministic $p \Rightarrow p(s,a,s') \in \{0,1\}$

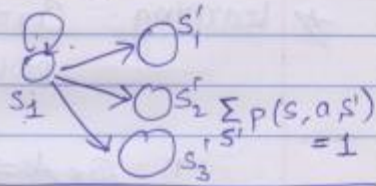
~~R: "Rewar~~ $R(20, \text{right}, 1) \rightarrow$ undefined

$R(20, \text{right}, 20) \Rightarrow 0$

$R(\underset{S_t}{20}, \underset{A_t}{\text{right}}, \underset{S_{t+1}}{21}) = -10 \rightarrow R_t$

What is $\sum_{s'} P(s,a,s')$ and $\sum_{s} P(s,a,s')$?





A note on R_t:

$R(\underset{S_t}{20}, \underset{a_t}{\text{right}}, \underset{S_{t+1}}{21}) = -10$ [We're already conditional on S_{t+1} (see formula for R_t) —

$R(20, \text{right}) = 0.8 \times -10 + 0.05 \times 0 + 0.5 \times 0 + 0.2 \times 0.1 \ldots$

used answering a question. Hereafter we use $R(s,a,s')$, not $R(s,a)$

Why take an expectation over R?

5. d_0 : Initial state distribution

$d_0 : S \rightarrow [0,1]$

$d_0(s) \triangleq \Pr(S_0 = s)$.

6. γ = Reward discount parameter

$\gamma \in [0,1]$.

Agent formulation (for an MDP)

* Policy: The mechanism within the agent that determines which action to take in a state.

* Learning: Corresponds to agent changing its policy.

$$\pi : S \times A \Rightarrow [0,1]$$
$$\pi(s,a) \triangleq Pr(A_t = a \mid S_t = s)$$

* Deterministic policies: A policy is deterministic if it always ~~tal~~ chooses same action in a given state.

$$\pi(s,a) \in \{0,1\} \quad \therefore \pi \text{ is deterministic}$$

* Stochastic policies:

π actions

	up	down	left	right
1	.01	.01	.08	.9
2	0	0	0.5	0.5
3				
4				
state 5				
⋮				
23				

All matrices with positive entries and rows summing to 1 are policies

(Representation of π)

$M = (S, A, P, R, d_0, \gamma)$

$S_0 \sim d_0$ (initial state sampled)

$a_0 \sim \pi(S_0, \cdot)$ (a_0 sampled)

$S_t \sim p(S_0, A_0, \cdot)$ notation to denote "any final state"

R_0 (random variable) $\rightarrow$ Computed with $\mathbb{E}[R_0] = R(S_0, A_0, S_1)$

↳ Deterministic reward

$R_0 = R(S_0, A_0, S_1)$.

Agent's goal:

Find a policy that maximizes the expected amount of reward that it will get.

Objective function: $J : \Pi \Rightarrow \mathbb{R}$ → set of all policies

$$J(\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} R_t \,\middle|\, \pi\right].$$

(This is a common objective function, but there are others.)

But a policy isn't an event so what does it mean to condition on π?

↳ $A_0 \sim \pi(s_0, \cdot)$
$A_1 \sim \pi(s_1, \cdot)$
$A_2 \sim \pi(s_2, \cdot)$

- We assume a fixed policy (π) ~~that d~~ across all time steps.

- This objective treats rewards as additive quantities, but other objective functions ~~could~~ may not.

Optimal ~~fo~~policy

$$\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$$

[argmax returns the set of optimal policies (there can be multiple)]

If $|S|$ and $|A|$ are finite, and R_t is bounded, then an optimal policy exists.