09/12 Lec - 3

# RL

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

$\pi(s, a) = Pr(A_t = a \mid S_t = s)$

$J(\pi) = E\left[\sum_{t=0}^{\infty} R_t \mid \pi\right]$

$p = 0.1$ lazy

$p = 0.8$ success

$p = 0.05$ +90

$p = 0.05$ −90

~~R(12)~~

$R(12, \text{attempt right}, 12) = E[R_t \mid S_t = 12, A_t = \text{Right}, S_{t+1} = 12]$

$= -10 \cdot \frac{0.8}{0.9}$

(after normalizing)

## Reward Discount : $\gamma \in [0,1]$

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

for $\gamma < 1$, rewards that occur later are worth less to the agent

→ initial ~~rwrd~~ reward is not discounted

$\gamma$ closer to 0 ≡ greedy

$\gamma$ closer to 1 ≡ long term planning

- including $\gamma$ ensures $J(\pi)$ is finite

  eg if $R_t \in [0,1]$

$$0 \le \sum_{0}^{\infty} \gamma^t R_t \le \frac{1}{1-\gamma}$$

- also helps ensures agent does what we want it to do

## Terminal States : A terminal state is a state which always transitions to a special state $S^\infty$ called the "terminal absorbing state"

1. "absorbing" - once there, the agent can ~~no~~ never leave

2 - transitioning into $S^\infty$, ~~the agent can never~~ gives 0 reward

3. All rewards in $S^\infty$ are zero

- "absorbing state" is part of the state space.

**Episode**: When an agent reaches $S^{\infty}$ the current trial called an "episode" ends and we start a new one

set $t \to 0$

and sample $s_0 \sim d_0$

→ Episodic MDP vs Continuous MDP

Agent always reaches $S^{\infty}$

---

**RL Definition revisited**

RL - requires "interaction" with the env, at least P is unknown

Planning: agent can solve w/o interacting the environment

---

- A **history**, $H_t$, is the history of what has happened up to time $t$

$$H_t \triangleq (S_0, A_0, R_0, S_1, A_1, R_1 \dots S_t, A_t, R_t)$$

- A **trajectory** is the history of an entire episode $= H_{\infty}$
- **Discounted Return**: of a trajectory is the discounted sum of rewards $G \triangleq \sum_{t=0}^{\infty} \gamma^t R_t$

$$J(\pi) = E[G|\pi]$$

# Mountain Car

**State:** $S = (X, V)$

**Actions:** $a \in \begin{cases} \text{forward, neutral, reverse} \\ (+1), \quad (0), \quad (-1) \end{cases}$

**Dynamics:**

$$V_{t+1} = V_t + 0.001a - [\cos(3x)](0.0025)$$

$$X_{t+1} = X_t + V_t$$

- if $X_{t+1}$ is outside $[-1.2, 0.5]$, it is moved to this range and $V_{t+1} \leftarrow 0$
- Terminal State: $X_t = 0.5$
- Rewards: $-1$ everywhere
- Initial State: $-0.5$
- Discount: $\gamma = 1$

Markov Problems: A seemingly more general formulation

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

Markov assumption / property:

$$Pr(S_{t+1} = s' \mid H_{t-1} = h, S_t = s, A_t = a) = Pr(S_{t+1} = s' \mid A_t = a, S_t = s)$$

or

The future is independent of the past given the present.

$$Pr(R_t = r \mid H_{t-1} = h, S_t = s, A_t = a, S_{t+1} = s')$$
$$= Pr(R_t = r \mid S_t = s, A_t = a, S_{t+1} = s')$$

Markovian Policy:

$$Pr(A_t = a \mid S_t = s, H_{t-1} = h) = Pr(A_t = a \mid S_t = s)$$

---

- Mountain Car:
  - → Markovian if $S = (X, V)$
  - → Not Markovian if $S = (X)$

⇒ Markovian is a property of the state representation.

---

One can always define a Markovian state representation given a non-Markovian state rep.

Let $S_t$ be a non-Markovian state rep

Then $(S_t, h_{t-1})$ is a Markovian state rep

but this might blow up the state space.