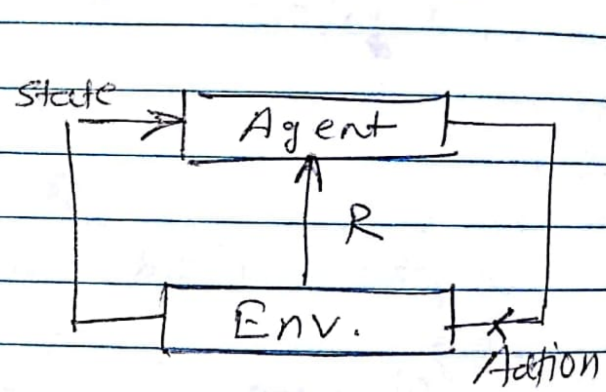
Lecture 5 ... 12th Sept 2017

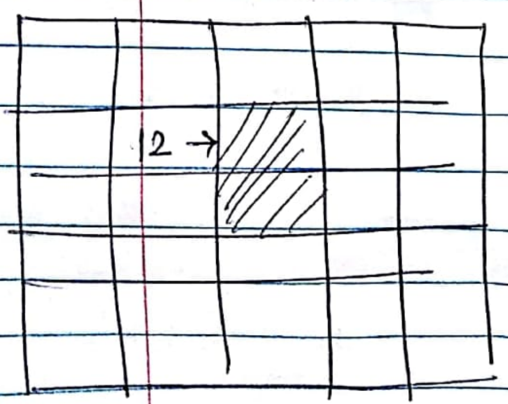
MDP $M = (S, A, P, R, d_0, \gamma)$

S = state,
A = action
P = transition f^n
R = reward f^n
d_0 = initial distribut^n
γ = discounting par.



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

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



$R(12, \text{attempt-right}, 12)$

$$= \mathbb{E}\left[R_t \mid S_t = 12, A_t = A\text{-}R, S_{t+1} = 12\right]$$

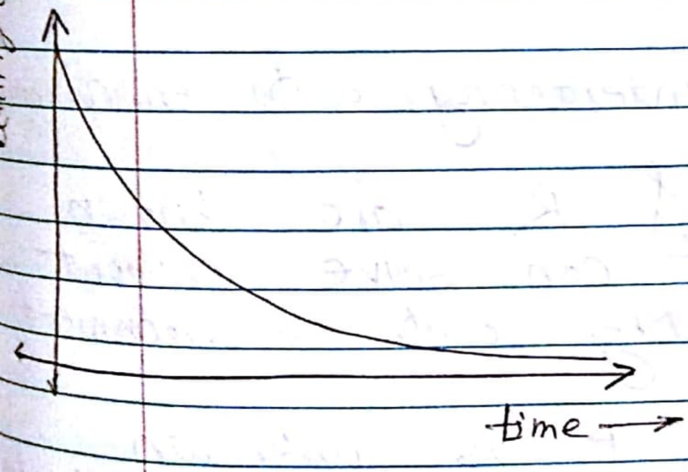
$$= \frac{-10 \times 0.8}{0.9}$$

Today's topic Environment

Reward discount (γ)

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

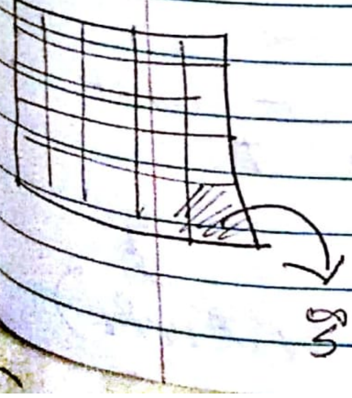
$$J(\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, \pi\right]$$



- Including γ (<1) ensures that $J(\pi)$ is finite

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

Terminal states: it is a state that always transition to a special state, called the 'Terminal absorbing state'



- Reward of transitioning into $\tilde{s}$ is zero
- Once enter $\tilde{s}$, can't get out

Episode : when agent reaches $\tilde{s}$, the current trial called an "episode", ends, and we start new one.

- set $t=0$
- sample $s_0 \sim d_0$

- Episodic MDP is one where agent always reaches $\tilde{s}$

RL Definition Revisited :

Learn from interacting with environment.

- Planning: P & R are known, agent can solve without interacting with environment

- RL : Atleast P is unknown, most methods will not estimate P.

Terminology :

① A history, H_t, is history of what has happened upto time t

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

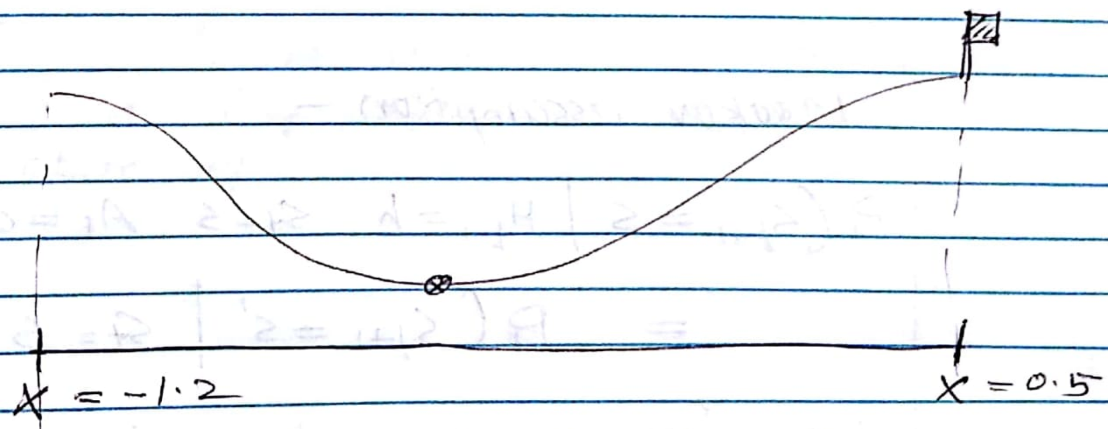
② <u>Trajectory</u> : is the history of an entire episode H_∞.

③ <u>Discounted Reward</u> : of trajectory is the discounted sum of rewards

$$G \triangleq \sum_{t=0}^{\infty} \gamma^{t} \cdot R_t$$

$$J(\pi) = \mathbb{E}[G \mid \pi]$$

<u>Mountain Car</u> :



State : $S = (x, v)$

Action : $a \in \{$forward, neutral, reverse$\}$

Dynamics : $v_{t+1} = v_t + 0.001a - 0.0025\cos(3x)$

Dynamics : $x_{t+1} = x_t + v_t$

if x_{t+1} is outside $[-1.2, 0.5]$ it is moved to this range & $v_{t+1} \leftarrow 0$

terminal state : $x_t = 0.5$
Reward : $R_t = -1$ everywhere
initial state :

Markov ~~Problem~~ Property

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

Markov assumption -

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

i.e. future is independent of past given the present

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

Code :: Blocks
Visual Studios

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

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

- Mountain car problem:

  if $s = (x, v)$ ✓
  $s = (x)$ ✗

Markovianess is a property of the state represertation

- One can always define a Markovian ~~st~~ state representat^n given a non-Markovian → —.

let S_t be a non-Markovian s. rep. then (S_t, h_{t-1}) is a Markovian state. rep.