Thetas[0] 1st sample from population

Lecture-6

Overview:
1. Environments
2. BBO for policy search
   - Simple agent that doesn't leverage MDP structure
3. Value functions
   - Tool for leveraging MDP
   - not a complete agent

State-Value function:

$$V^{\pi} : S \Rightarrow \mathbb{R}$$

$$V^{\pi}(s) \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^{k} R_{t+k} \,\middle|\, S_t = s, \pi\right]$$

$$\gamma^0 R_t + \gamma^1 R_{t+1} + \gamma^2 R_{t+2} \cdots$$

- Depends on policy π

= Expected discounted return from state s under policy π

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} R_{t+1} \,\middle|\, \pi\right]$$

$$= \sum_{s} d_0(s)\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} R_{t+k} \,\middle|\, \pi, S_0 = s\right]$$

$$\therefore J(\pi) = \sum_s d_0(s)\, V^{\pi}(s)$$

$$J(\pi) = V^{\pi}(s_0)$$

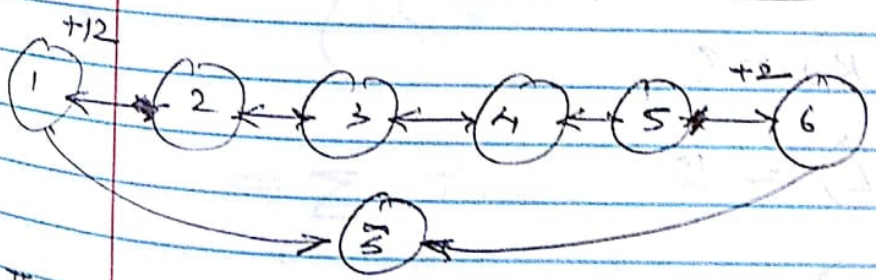
$$V^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\left[R_{t+k} \mid S_t = s, \pi\right]$$

$$= \gamma^0 \sum_a \pi(s,a) \sum_{s'} P(a,a,s') R(s,a,s')$$

$$+ \gamma^1 \sum_a \pi(s,a) \sum_{s'} P(s,a,s') \sum_{a'} \pi(s,a') \sum_{s''} P(s',a',s'') R(\cdot)$$

$$+ \ldots\ldots$$

Example on Value function



γ : 0.5

π_1 : left always

π_2 : Right always

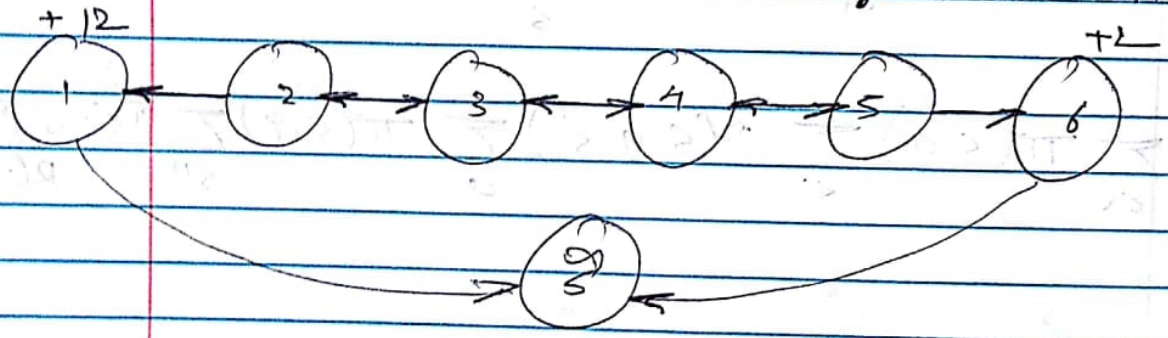
$V^{\pi_1}(s_1) = 0$	$V^{\pi_2}(s_1) = 0$
$V^{\pi_1}(s_2) = 12\gamma^0$	$V^{\pi_2}(s_2) = 1/4$
$V^{\pi_1}(s_3) = 12\gamma^1$	$V^{\pi_2}(s_3) = 1/2$
$V^{\pi_1}(s_4) = 12\gamma^2$	$V^{\pi_2}(s_4) = 1$
$V^{\pi_1}(s_5) = 1.5$	$V^{\pi_2}(s_5) = 2$
$V^{\pi_1}(s_6) = 0$	$V^{\pi_2}(s_1) = 0$

Action value function / state-action value f^n / q-f^n

$q^{\pi} : S \times A \Rightarrow \mathbb{R}$

$$q^{\pi}(s,a) \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s, A_t = a, \pi \right]$$

= Expected discounted return from state s if the agent takes action a & follows π, thereafter

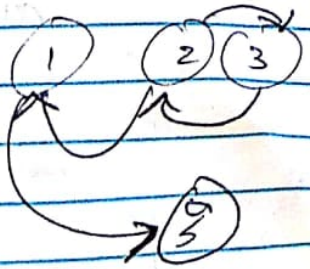


$q^{\pi_1}(s,L) = 0$

$q^{\pi_1}(s_1, R) = 0$

$q^{\pi_1}(s_2, L) = 12$

$q^{\pi_1}(s_2, R) = 3$



Bellman Equation for V^π

$$V^\pi(s) \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s, \pi\right]$$

$$= \mathbb{E}\left[\gamma^0 R_t + \sum_{k=1}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s, \pi\right]$$

$$= \mathbb{E}\left[\gamma^0 R_t + \sum_{k=0}^{\infty} \gamma^{k+1} R_{t+k+1} \,\middle|\, S_t = s, \pi\right]$$

$$= \mathbb{E}\left[R_t + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, \pi\right]$$

$$= \sum_a \pi(s,a) \sum_{s'} P(s,a,s') \times$$

$$\left(R(s,a,s') + \gamma\, \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, \pi\right]\right)$$

$$= \sum_a \pi(s,a) \sum_{s'} P(s,a,s') \left[R(s,a,s') + \gamma V^\pi(s')\right]$$

$$= \mathbb{E}\left[R_t + \gamma V^\pi(S_{t+1}) \,\middle|\, S_t = s, \pi\right]$$

$$= R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} \cdots$$