Bellman Equations

Prof. Subrahmanya Swamy

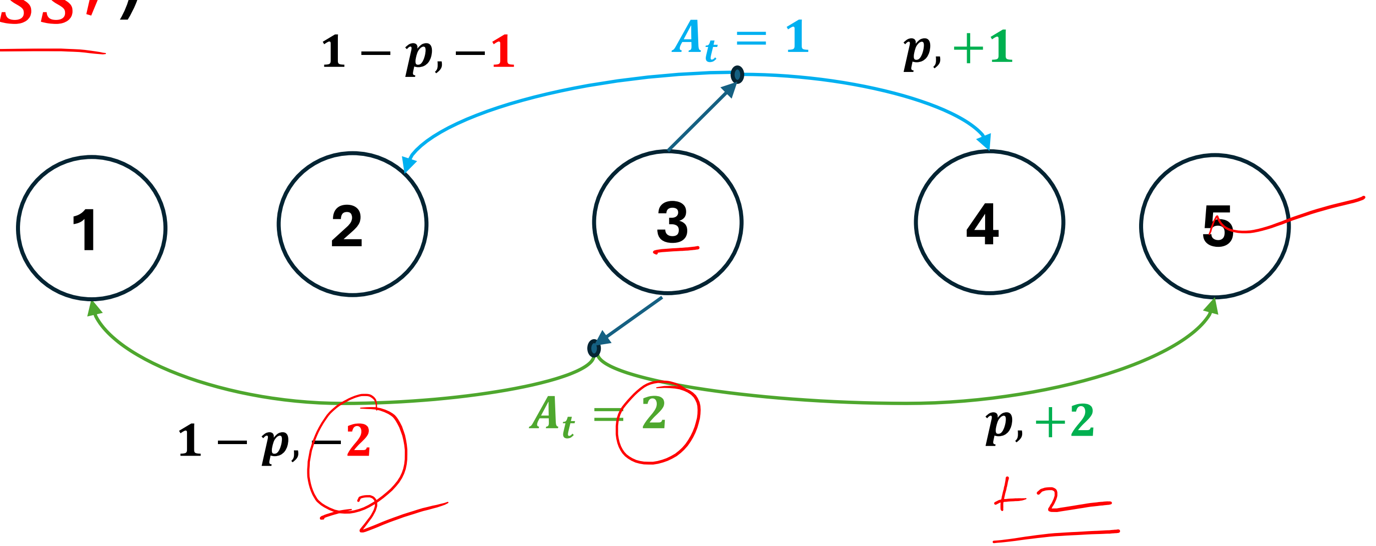
Outline

- MDP Dynamics $R_s^a, P_{ss'}^a$
- Policy Dynamics $R_s^\pi, P_{ss'}^\pi$
- Value Function $V_\pi(s)$
- Action-Value Function $Q_\pi(s, a)$
- Bellman Equations

MDP Dynamics ($R_s^a, P_{ss'}^a$)

Transition Probability

- $P_{ss'}^a = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$
- *Example:* $P_{3,5}^2 = p$ ✓



Expected Reward

- $R_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- *Example:*

$$R_3^2 = 2p - 2(1-p)$$
$$= 4p - 2$$
$$= -1 \quad (if\ p = \tfrac{1}{4})$$

Policy Dynamics $(R_s^\pi, P_{ss'}^\pi)$

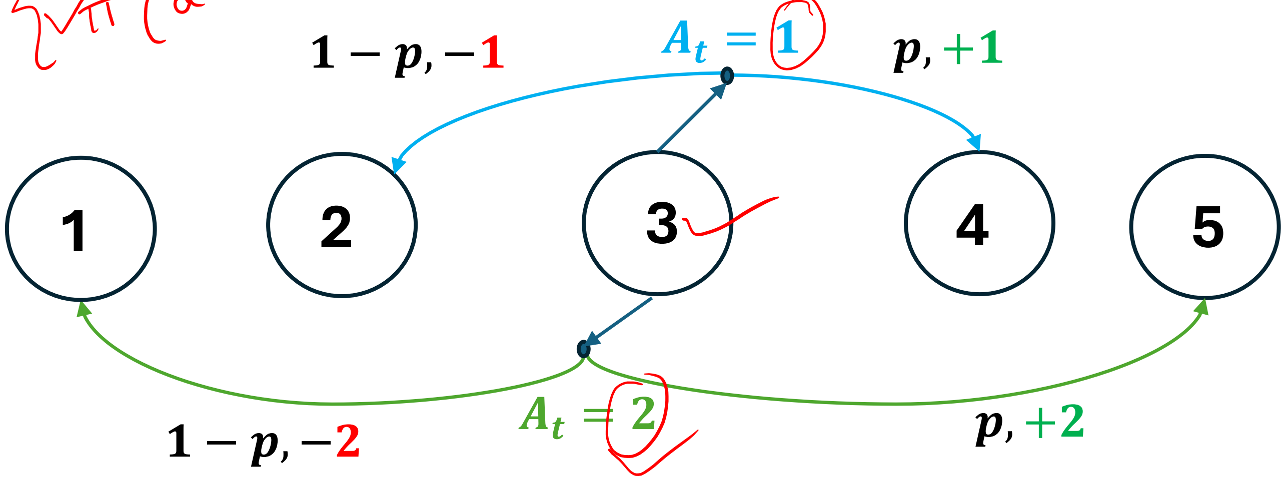
Transition Probability

- $P_{ss'}^\pi = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t \sim \pi)$
- $= \sum_a \pi(a \mid s)\, P_{ss'}^a$

$\pi(a=2 \mid s=3) = 1/2$

$\pi(a=1 \mid s=3) = 1/2$

$P_{ss'}^\pi = \frac{1}{2} \cdot P_{ss'}^2 + \frac{1}{2} P_{ss'}^1$



Expected Reward

- $R_s^\pi = \mathbb{E}[R_{t+1} \mid S_t = s, A_t \sim \pi]$
- $= \sum_a \pi(a \mid s)\, R_{ss'}^a$

R_3^π

Value Function ($V_\pi(s)$)

Bellman Expectation Eq.

The expected return for following policy π starting from state s

$\pi(\text{Left}|s) = 1/4$

$\pi(R|s) = 1/4$

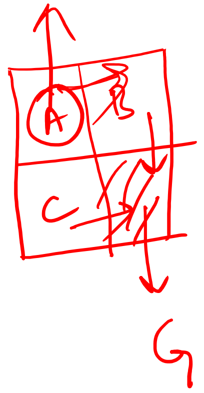
$\pi(D|s) = 1/4$

$\pi(U|s) = 1/4$

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

s = A

$V_\pi(A)$



$P(A, \text{Right}, -1, B, \text{Down}, -1, G) \rightarrow -2$

$A, \text{UP}, -1, A, \text{Down}, -1, C, \text{Right}, -1, G \rightarrow -3$

A A A A A ... B G

$\pi(\text{Right}) \times \pi(\text{Down}) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$

$-2 \times \frac{1}{16} \quad -3 \times \left(\frac{1}{4}\right)^3 =$

Action-Value Function ($Q_\pi(s,a)$)

The expected return for taking action a in current state s and then following policy π from the next state

$$Q_\pi(s,a) \coloneqq \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$V_\pi(s)$ $Q_\pi(s,a)$

Relating Q_π and V_π

$$V_\pi(s) = \sum_a \pi(a \mid s)\, Q_\pi(s, a)$$

$\pi \rightarrow a_1 \rightarrow \pi(a_1|s)$

$\quad\ \rightarrow a_2 \rightarrow \pi(a_2|s)$

$V_\pi(s) = \pi(a_1|s) Q_\pi(s, a_1) + \pi(a_2|s) Q_\pi(s, a_2)$

$Q_\pi(s, a)$ $\quad V_\pi(s)$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} \cdots$

$$Q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$= \underbrace{E_\pi\left[R_{t+1} \mid S_t = s, A_t = a\right]}_{R_s^a} + \gamma\, \underline{E_\pi\left[G_{t+1} \mid S_t = s, A_t = a\right]}$$

$$= R_s^a + \gamma$$

$E[X+Y]$
$= E[X] + E[Y]$

Relating Q_π and V_π

$V_\pi \leftrightarrow Q_\pi$

$V_\pi (A)$
$V_\pi (B)$
$V_\pi (C)$

- $$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$
  $$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$
  $$= \mathbb{E}_\pi[R_{t+1} \mid S_t = s, A_t = a] + \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a]$$
  $$= R_s^a + \gamma \sum_{s'} P_{ss'}^a \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s', \cancel{S_t = s}, \cancel{A_t = a}]$$
  $$Q_\pi(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s') \quad (1)$$

Markov property

- Substitute this in $V_\pi(s) = \sum_a \pi(a \mid s) \boxed{Q_\pi(s,a)}$ to get V_π interms of V_π

(2)

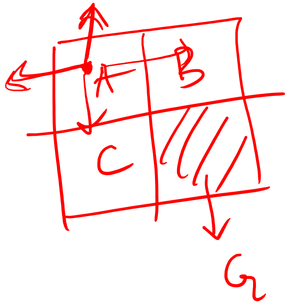
Bellman Expectation (BE) equation

- V_π in terms of V_π : (Useful to compute V_π from $P_{ss'}^{a}$ and R_s^a)

$$V_\pi(s) = \underbrace{R_s^\pi}_{\text{Immediate reward}} + \sum_{s'} P_{ss'}^{\pi}\, \underbrace{V_\pi(s')}_{\text{Remaining Return}}$$

- $R_s^\pi := \sum_a R_s^a\, \pi(a \mid s)$
- $P_{ss'}^\pi := \sum_a P_{ss'}^a\, \pi(a \mid s)$

Example



$\pi \rightarrow$ uniform Random Policy

BE

$$V_\pi(s) = \underbrace{R_s^\pi}_{①} + \underbrace{\gamma \sum_{s'} P^\pi_{ss'} V_\pi(s')}_{②}$$

① Immediate Reward

② Remaining Return

$10^6 \times 10^6$

3 unknowns

3 linear equations

$V_\pi(A)$

① $R^\pi_A = -1$

② $P^\pi_{AB} = \frac{1}{4}$, $P^\pi_{AC} = \frac{1}{4}$, $P^\pi_{A,A} = \frac{1}{2}$

$Ax = b$

$\boxed{x = A^{-1}b}$

①
$$\begin{cases} V_\pi(A) = -1 + \gamma\left(P^\pi_{AB} V_\pi(B) + P^\pi_{AC} V_\pi(C) + P^\pi_{AA} V_\pi(A)\right) \\ V_\pi(B) = -1 + \ldots \\ V_\pi(C) = -1 + \ldots \end{cases}$$