Bellman Equations

Prof. Subrahmanya Swamy

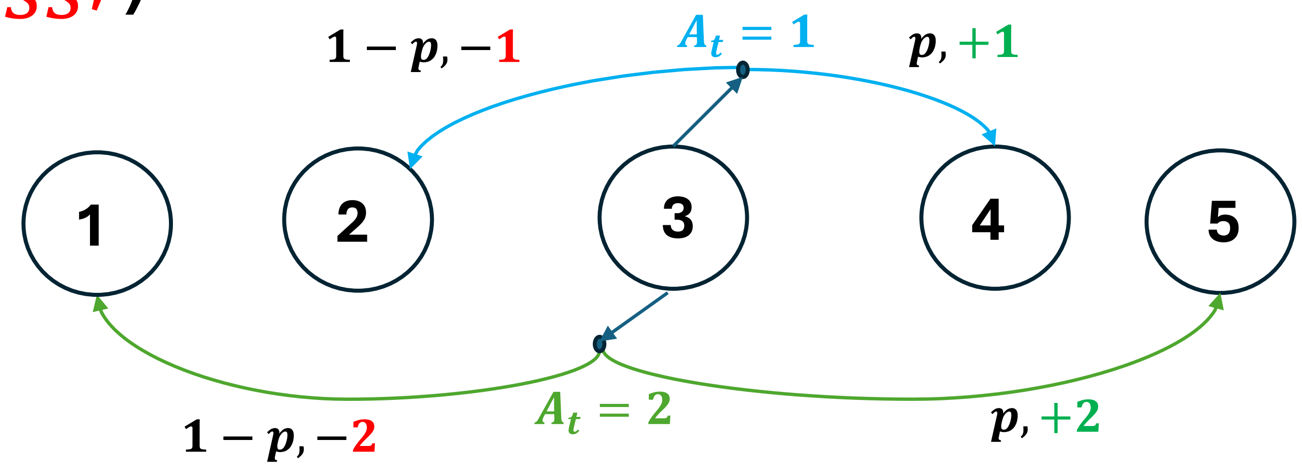
Outline

- 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

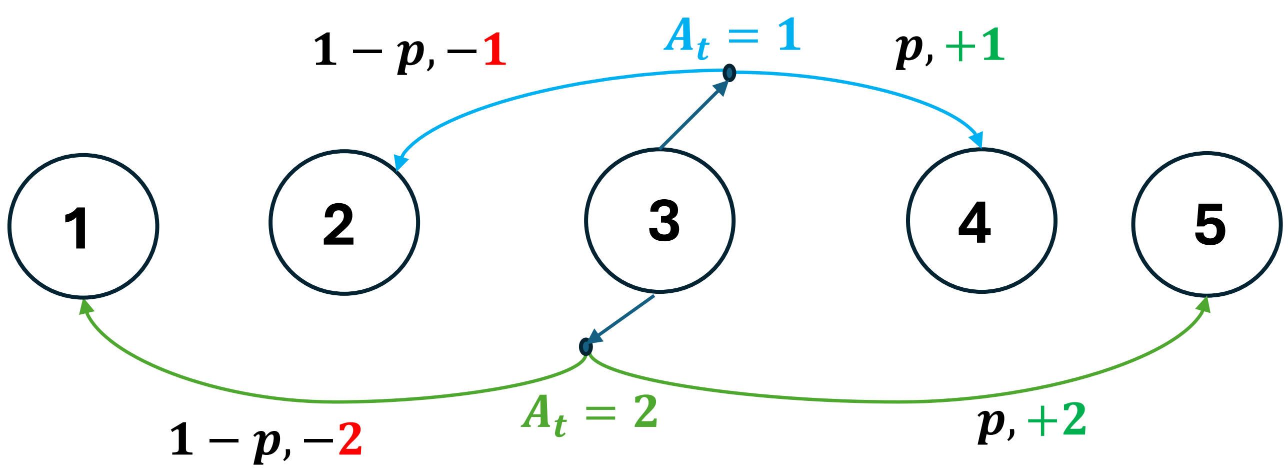
- $R_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- *Example:*

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

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

Transition Probability

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



Expected Reward

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

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

The expected return for following policy π starting from state s

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

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

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

Relating Q_π and V_π

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

Relating Q_π and V_π

- $$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + G_{t+1} \mid S_t = \mathrm{s}, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} \mid S_t = \mathrm{s}, \mathrm{A_t} = \mathrm{a}] + \mathbb{E}_\pi[G_{t+1} \mid S_t = \mathrm{s}, \mathrm{A_t} = \mathrm{a}] \\ &= R_s^a + \textstyle\sum_{s'} P_{ss'}^a \, \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s',\ S_t = s,\ A_t = a] \\ &= R_s^a + \textstyle\sum_{s'} P_{ss'}^a \, V_\pi(s') \end{aligned}$$

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

Bellman Expectation (BE) equation

- V_π in terms of V_π : (Useful to compute V_π from $P^a_{ss'}$ and R^a_s)

$$V_\pi(s) = \underbrace{R^\pi_s}_{\text{Immediate reward}} + \sum_{s} P^\pi_{ss'} \underbrace{V_\pi(s')}_{\text{Remaining Return}}$$

- $R^\pi_s := \sum_a R^a_s\, \pi(a \mid s)$
- $P^\pi_{ss'} := \sum_a P^a_{ss'}\, \pi(a \mid s)$