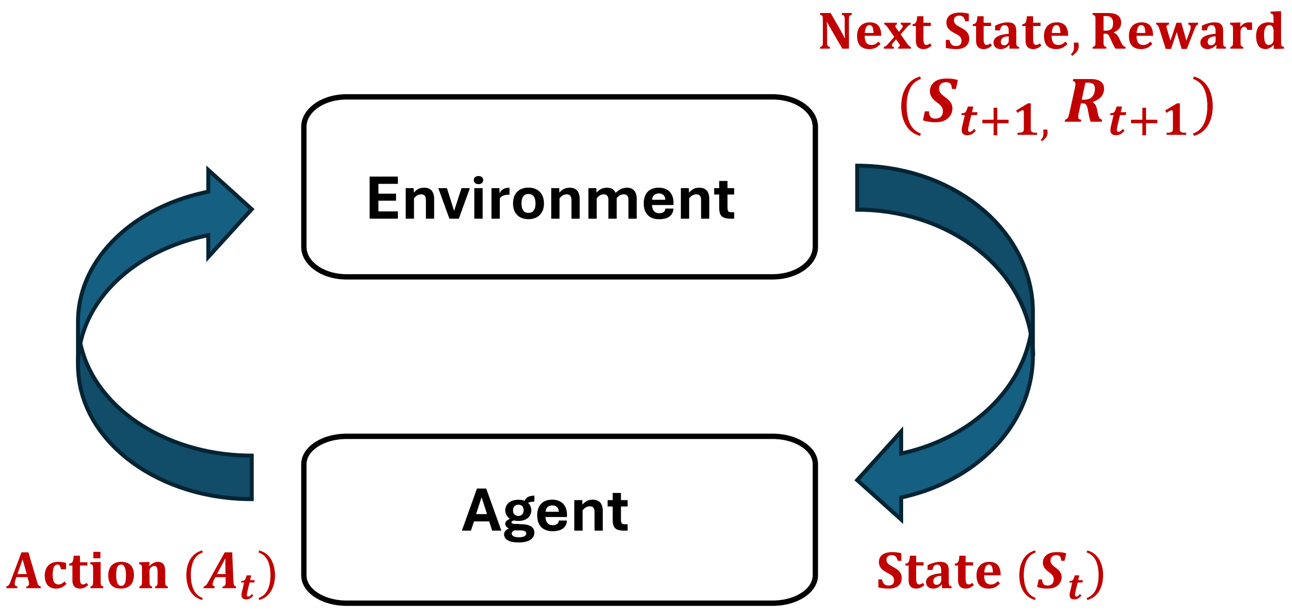
Markov Decision Processes (MDP)

Prof. Subrahmanya Swamy

RL Framework



1. Agent observes the state
2. Takes an action
3. Environment puts the agent in a new state &
4. Also gives a reward based on taken action

Goal:
Learn policy to maximize the cumulative reward $\sum_t R_t$

How do we mathematically model the "State transitions" and Rewards?

Independent Random Variables

- A sequence of coin tosses $X_1, X_2, X_3, \ldots$

1/0

- Head: 1, Tail: 0, Bias of coin: p_h
- Knowledge of X_1 does not help in predicting X_2
- $\mathbb{P}(X_2 = 1 \,|X_1 = 0) = p_h$
- $\mathbb{P}(X_2 = 1 \,|X_1 = 1) = p_h$

Markov Chain

- A sequence of coin tosses $X_1, X_2, X_3, \ldots$
- If coin lands in
  - Head: Win 1 rupee
  - Tail: Lose 1 rupee
- Define $Y_t =$ total money accumulated till time t
- $Y_1, Y_2, Y_3, \ldots$ are dependant RVs
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 3) = 0$
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 0) = \frac{1}{2}$

$Y_1 \quad Y_2$

$4 \leftrightarrow 2$

$Y_{t+1} = Y_t + X_t$

$+1 / -1$

Markov Chain

$P(Y_5 | Y_3, Y_4)$

$P(Y_5 | Y_4, Y_3, Y_2, Y_1)$

$\parallel$

$P(Y_5 | Y_4)$

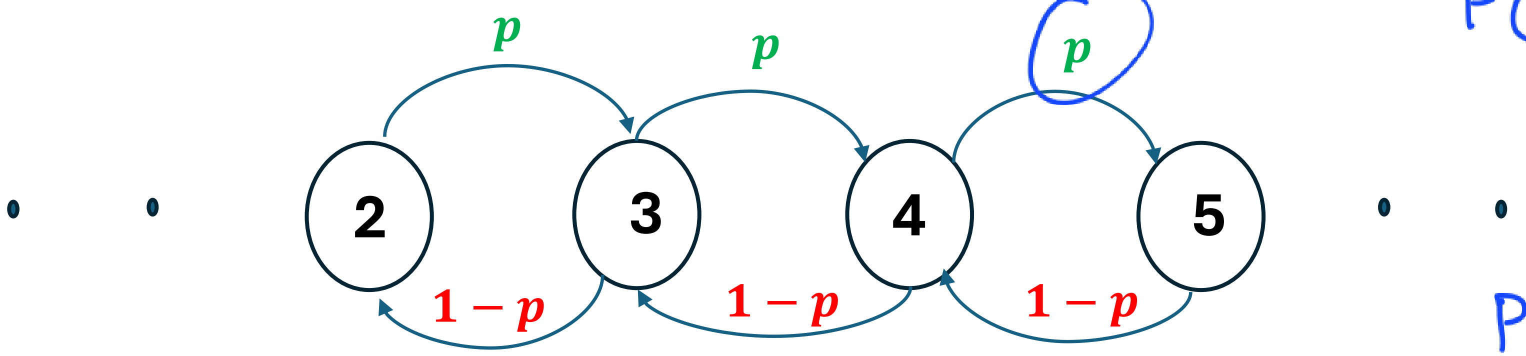
- $Y_1, Y_2, Y_3, \ldots$ satisfy Markov property!

- **Markov Property:** Given the present, the future is independent of the past!
  - $\mathbb{P}(Y_5 = 1 \,| Y_4 = 2, Y_3 = 3) = \frac{1}{2}$
  - $\mathbb{P}(Y_5 = 1 \,| Y_4 = 2, Y_3 = 1) = \frac{1}{2}$

+1 — H 1/2

−1 — T 1/2

$P(Y_5 = 0 | Y_4 = 2) = 0$



$P = 1/2$

3 = P(H) p

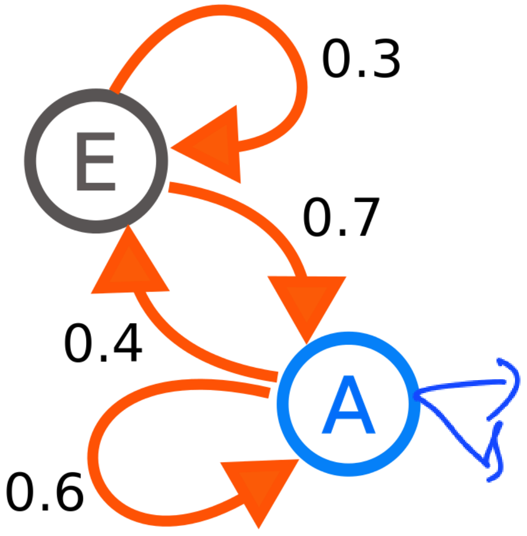
1 = P(T) 1−P

Markov Chain Specification $(S, P_{ss'})$

- $S \rightarrow State\ space \quad \{E, A\}$

- $P_{ss'} \rightarrow Transition\ probabilty$
  - $\mathbb{P}(S_{t+1} = s' \mid S_t = s)$

$$\begin{array}{ccc} & E & A \\ E & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$}} \\ A & & \end{array}$$

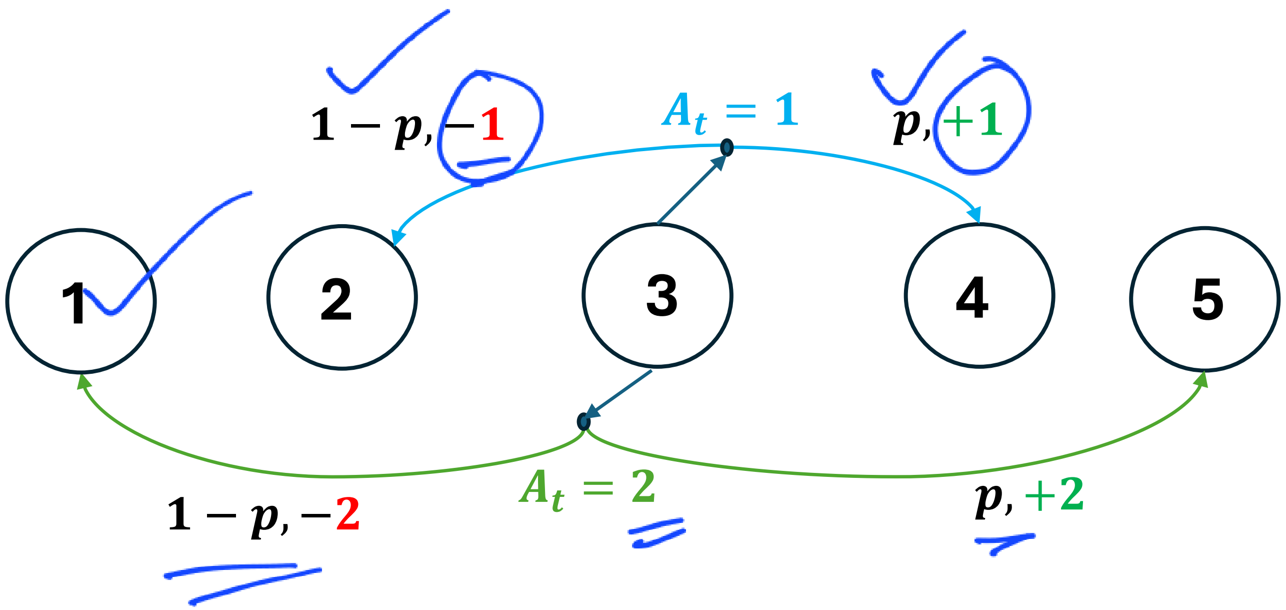


Markov Decision Process (MDP)

- Introduce action to convert Markov Chain into MDP

- Actions: How much money to bet (A_t) in the game when I have Y_t money?

- If $Y_t = 3$, then possible actions are $\{1,2,3\}$.



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

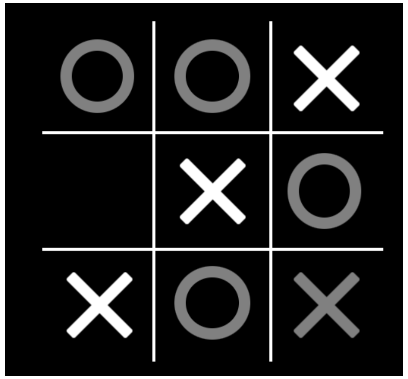
$$P(S_{t+1} = s' \mid S_t = s, A_t = a)$$

Episodic and Continuing MDPs

$S_0, A_0, R_1, S_1, A_1, \ldots, S_T$

- Episodic
  - There exists a special state called the terminal state
  - The episode ends at the terminal state
  - Eg: Board games

Terminal state in Tic-Tac-Toe



- "Continuous"
  - No terminal state exists
  - The task continues forever
  - Eg: Portfolio management
    - Every day, decide which shares to buy/sell

$S_0, A_0, \ldots\ldots$

AMC

Discount Factor in MDP

- Episodic task:
- Total Reward (Return) : $G_t = R_{t+1} + R_{t+2} + \ldots + R_T$
- Bounded Returns if each $R_i \leq M$

- Continuing task:
- $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots$
- $G_t = \sum_{i=t+1}^{\infty} R_i$ could become unbounded even if each $R_i \leq M$

- Solution: Discount factor $\gamma \in (0,1)$
- $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$
- $G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \leq \frac{M}{1-\gamma}$ (Bounded)

$\frac{1}{2}\left(\frac{1}{4}\right)^2$

$\leq M + \gamma M + \gamma^2 M + \ldots \quad \frac{M}{1-\gamma}$

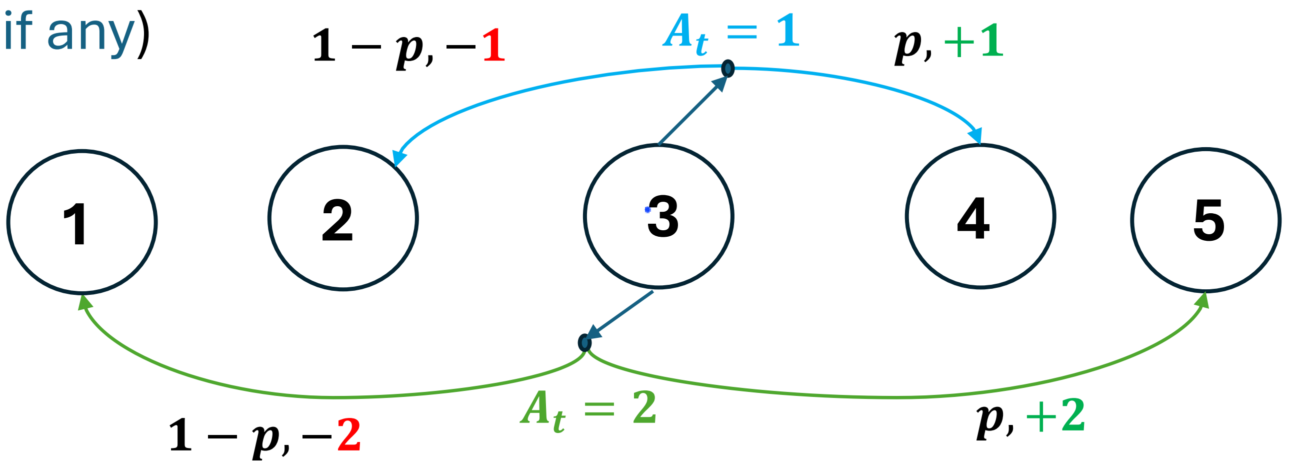
0.99

- High $\gamma \sim 1 \Rightarrow$ Long-term planning
- Low $\gamma \sim 0 \Rightarrow$ Short-term planning

MDP Specification $(S, A, R_s^a, P_{ss'}^a, \gamma)$

- $S \rightarrow State\ space$ (incl. terminal states if any)
- $A \rightarrow Action\ space$
- $R_s^a \rightarrow Expected\ Rewards$
- $\mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- $P_{ss'}^a \rightarrow Transition\ probabilties$
- $\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$
- $\gamma \in (0,1) \rightarrow Discount\ factor$

(s, s')



$5 \times 5 = 25$

$P^1_{3,4} = P$

$P^2_{3,1} = 1 - P$

$P^2_{3,5} = P$

$P^2_{3,6} = 0$

Optimal Policy

$\pi: S \to A$

- Policy:
  - Deterministic: $\pi(s): \mathcal{S} \to \mathcal{A}$ Which action to take in state s
  - Stochastic: $\pi(a \mid s)$ In state s, with what probability to take action a

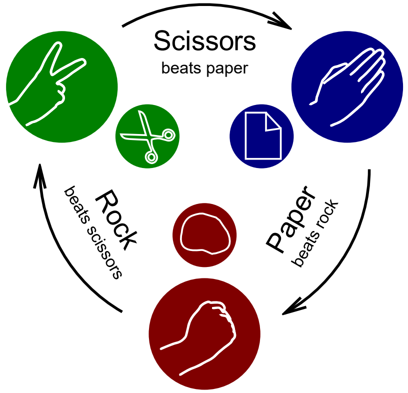
$3 \to 0 \to 1/4$

$1 \to 1/2$

$2 \to 1/8$

$3 \to 1/8$

- Why stochastic policies?
  - Partially observed states
  - Exploration



$\mu($

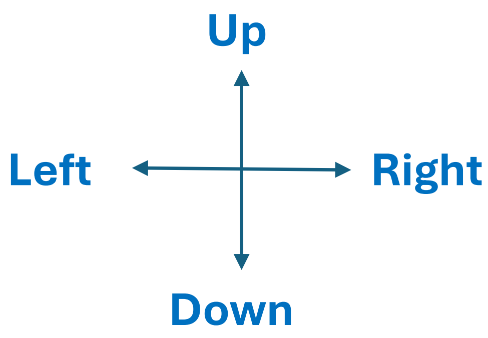
$1-\varepsilon$ exploit

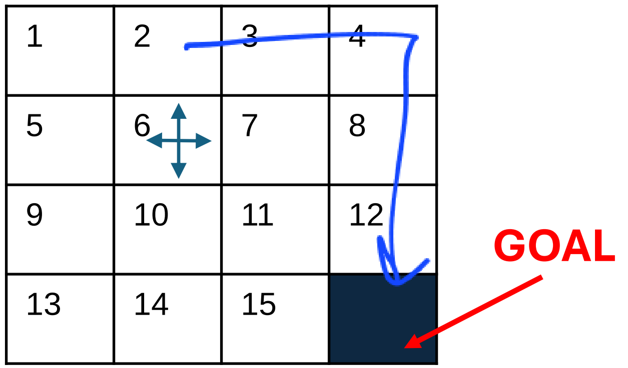
ε explore

- Optimal Policy:
  - π that maximizes the expected return $\mathbb{E}_\pi[G_t \mid S_t = s]$ from any state s

How to model your problem as an MDP?

Maze Solving Problem: To reach the goal in the shortest path!

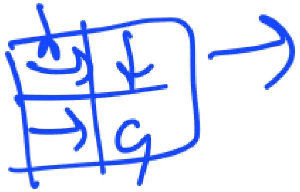


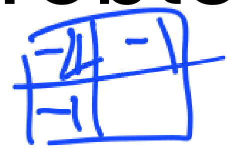


- How to formulate this maze-solving problem as an MDP?
  - States ?
  - Actions ?
  - Rewards ?
  - Transition Probabilities ?
  - Discount factor ?

S
A
R_s^a
$P_{ss'}^a$
γ .

How to model your problem as an MDP?





4 Actions

Up

Left — Right

Down

16 States

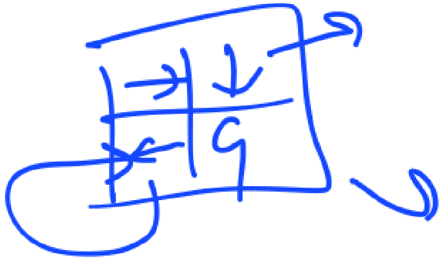
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

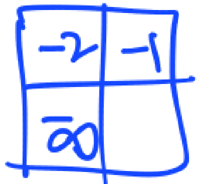
Terminal State

+1

−1

+1





Rewards

$\boldsymbol{R_t = -1}$ on all transitions

Discount Factor

$\boldsymbol{\gamma = 1}$

$\pi(s) \to a.$

Deterministic State transitions : $\mathbb{P}\,(S_{t+1} = 2 \mid S_t = 6, A_t = Up) = \mathbf{1}$

Verify that optimal policy = shortest path

$P^R_{3,4} = 1$

$G_t = R_{t+1} + \gamma R_{t+2} \ldots \gamma^{t} R_T$

Exercise

- Alternate MDP formulation for the Maze problem

- Instead of giving -1 reward per each step, can we give 0 reward for every action except for the final action that leads us to the Goal State?

+1

0 . . . . +1

- Does the optimal policy of this alternate MDP learn the shortest path?

- Hint: What discount factor will help here?

$\gamma < 1$

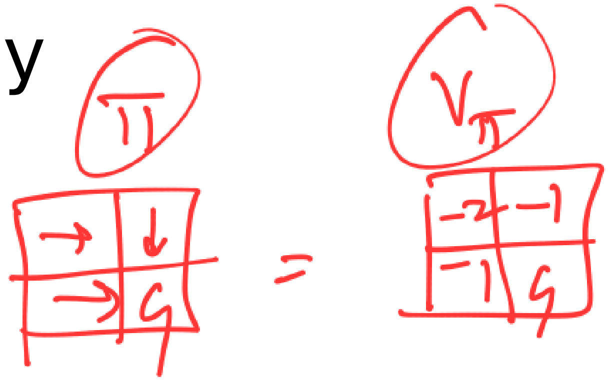
$\gamma = 0.9$

$0 + \gamma \cdot 1 = 0.9$ ✓

$\rightarrow G_t = 0 + \gamma \cdot 0 + \gamma^2 \cdot 1 = 0.81$ ✓ . . .

Bellman Equations

Prof. Subrahmanya Swamy



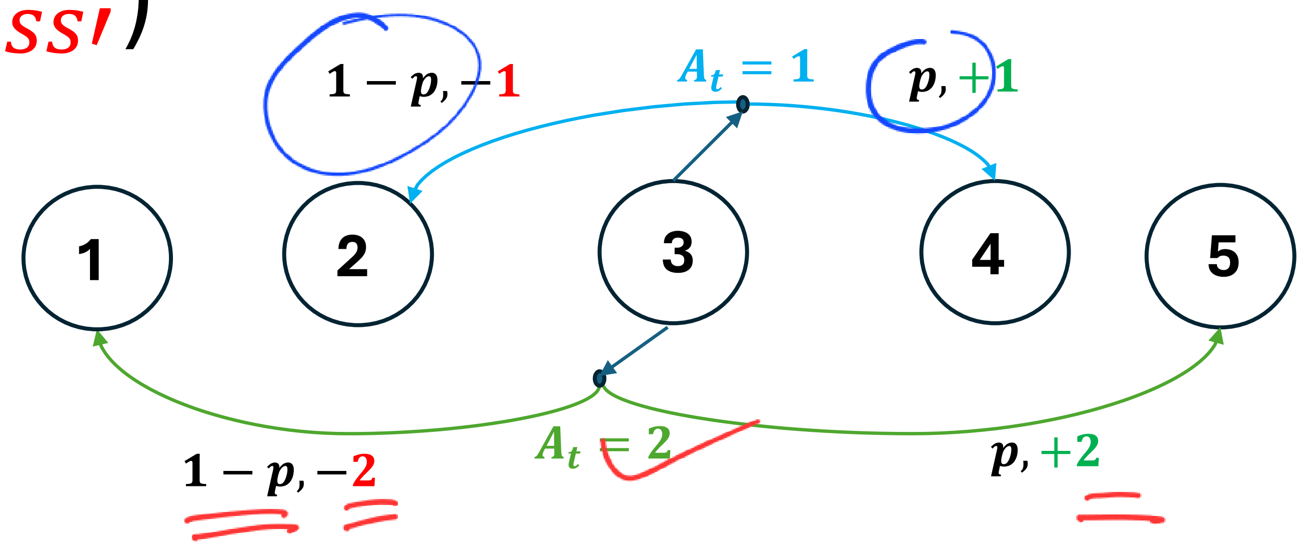
Outline

- MDP Dynamics $\boxed{R_s^a, P_{ss'}^a}$
- 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 $(R_s^\pi, P_{ss'}^\pi)$

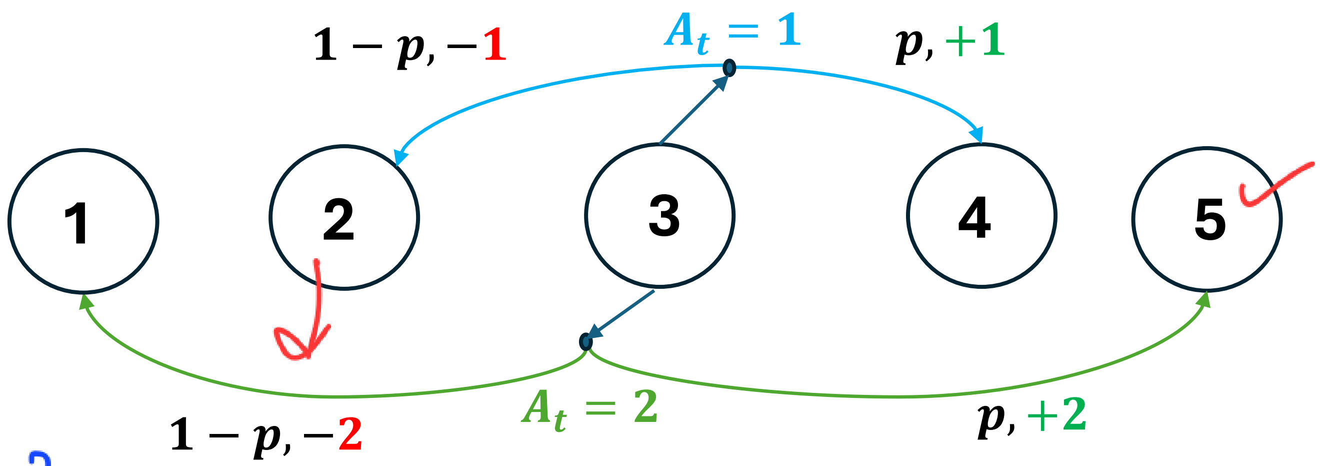
Transition Probability

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

$\frac{1}{4} P^1_{ss'} + \frac{1}{8} P^2_{ss'} + \frac{1}{8} P^3_{ss'} + \frac{1}{2} P^0_{ss'}$

Expected Reward

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



$\pi(a=1 \mid s=2) = 1/4$

$\pi(a=2 \mid s=2) = 1/8$

$3 \quad = 1/8$

$0 \quad = 1/2$

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

The expected return for following policy π starting from state s

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

"Bellman Eq."

"−1"

$\gamma = 1$

π

$-1 \times 1/4$

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

-5

1/4

1/4 1/4

$B \to G \to -1$

B A C A B G

$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} =$

$\left(\frac{1}{4}\right)^5$

$$\sum_{\text{paths}} P(\text{path})\, G_t(\text{path})$$

$$E[X] = \sum_x P(x) \cdot x$$

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

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

B

L

π 1/4 1/4 1/4 1/4

B

V_π Q_π

Bellman