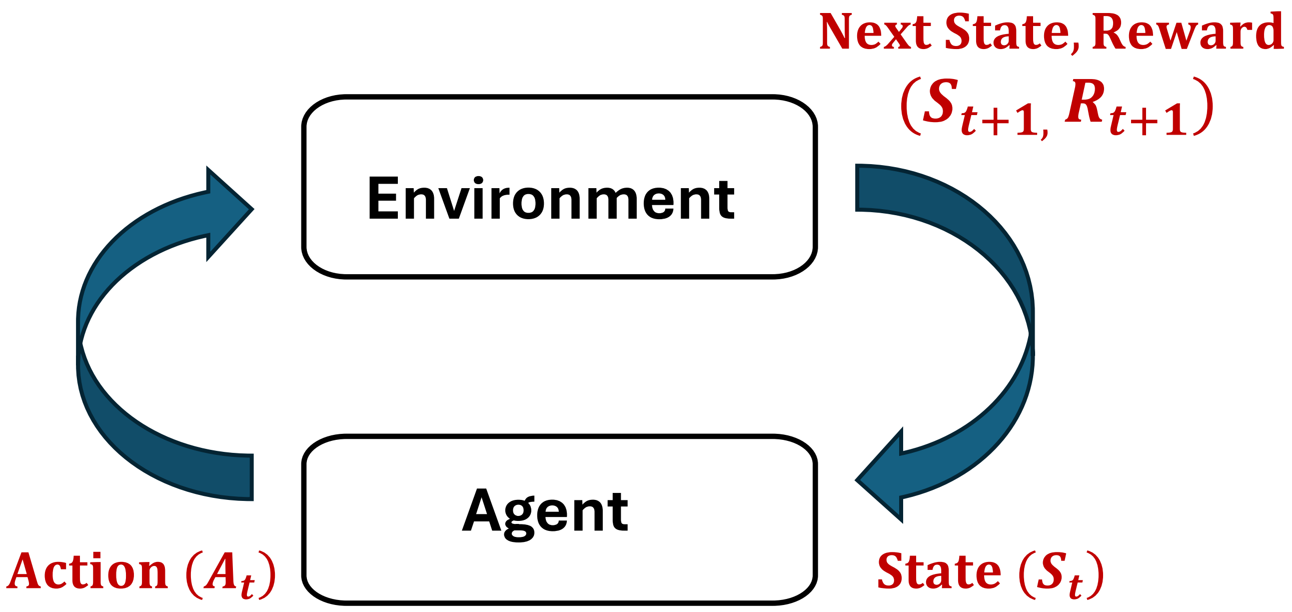
Markov Decision Processes (MDP)

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 $X_1, X_2, X_3, \ldots$

- Head: 1,   Tail: 0,   Bias of coin: p_h

- Knowledge of X_1 does not help in predicting X_2

- $\mathbb{P}(X_2 = 1 \mid X_1 = 0) = p_h$
- $\mathbb{P}(X_2 = 1 \mid X_1 = 1) = p_h$

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

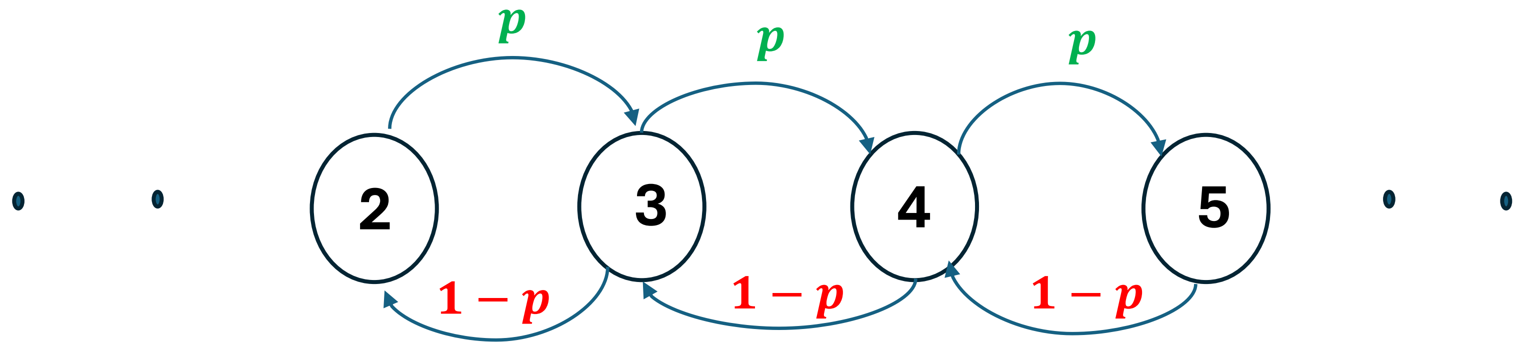
- Define Y_t = total money accumulated till time t

- $Y_1, Y_2, Y_3, \ldots$ are dependant RVs
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 3) = 0$
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 0) = \frac{1}{2}$

Markov Chain

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

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

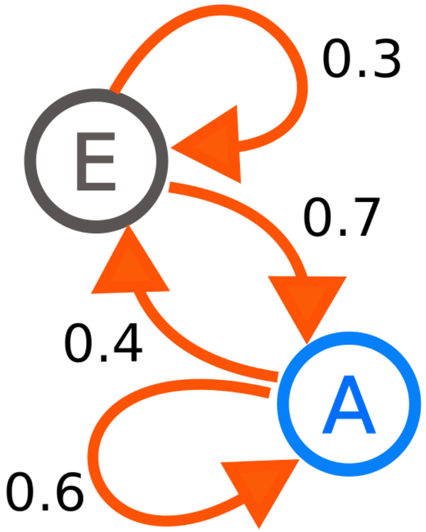


Markov Chain Specification $(S, P_{ss'})$

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

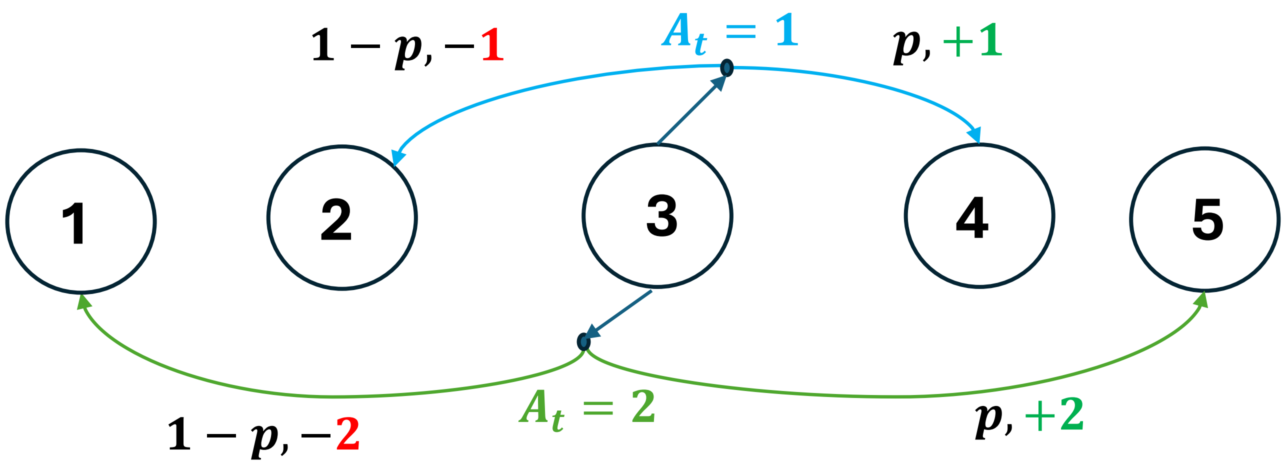
- $P_{ss'} \;\rightarrow Transition\ probabilty$
  - $\mathbb{P}(S_{t+1} = s' \mid S_t = s)$

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



Markov Decision Process (MDP)

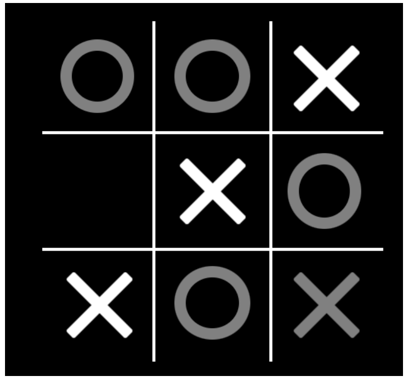
- Introduce action to convert Markov Chain into MDP
- Actions: How much money to bet (A_t) in the game when I have Y_t money?
- If $Y_t = 3$, then possible actions are $\{1,2,3\}$.



Episodic and Continuing MDPs

- 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

Discount Factor in MDP

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

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

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

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

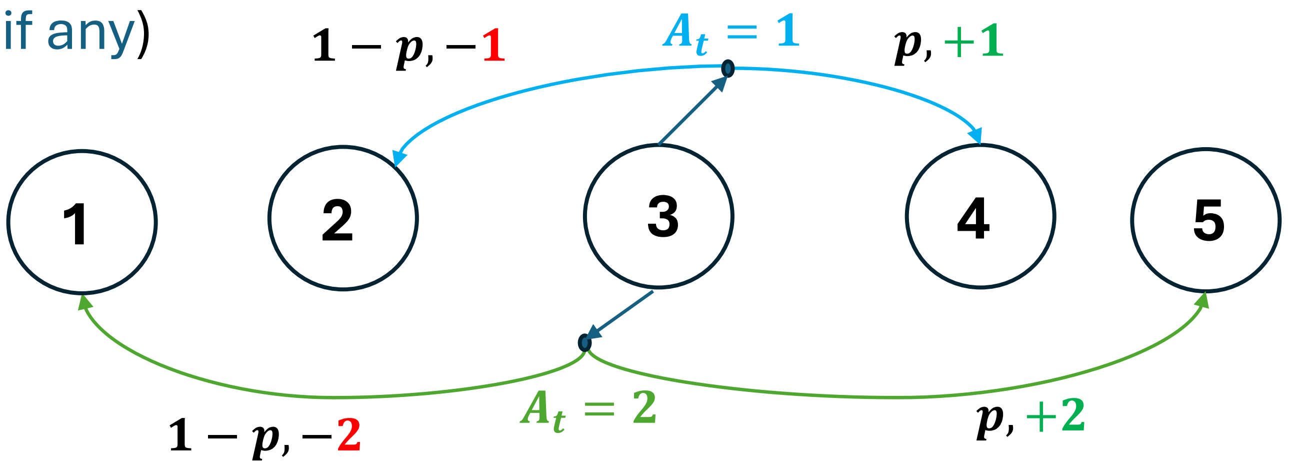
MDP Specification $(S, A, R_s^a, P_{ss'}^a, \gamma)$

- $S \rightarrow State\ space$ (incl. terminal states if any)
- $A \rightarrow Action\ space$

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

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

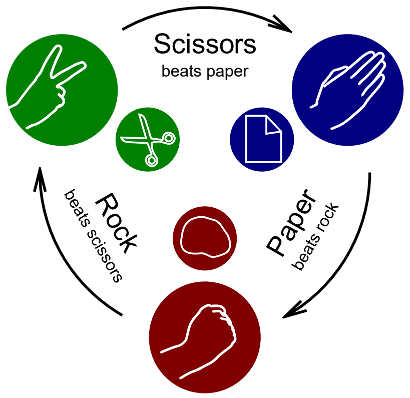
- $\gamma \in (0,1) \rightarrow Discount\ factor$



Optimal Policy

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

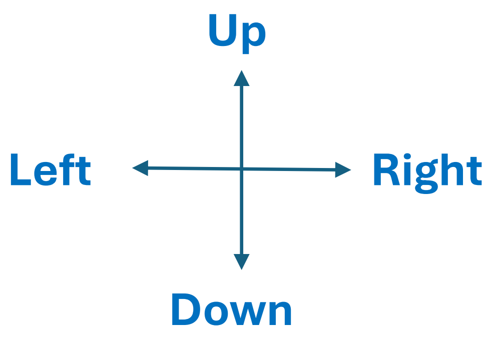
- Why stochastic policies?
  - Partially observed states
  - Exploration

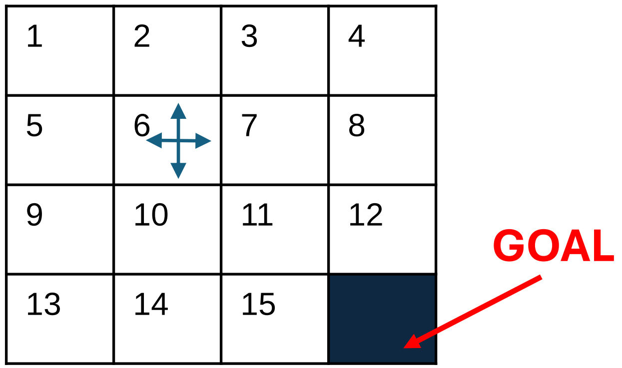


- Optimal Policy:
  - π that maximizes the expected return $\mathbb{E}_\pi[G_t \mid S_t = s]$ from any state s

How to model your problem as an MDP?

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

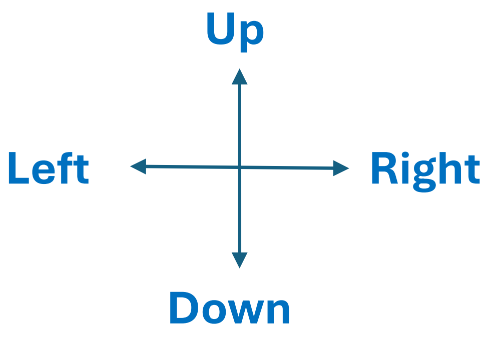




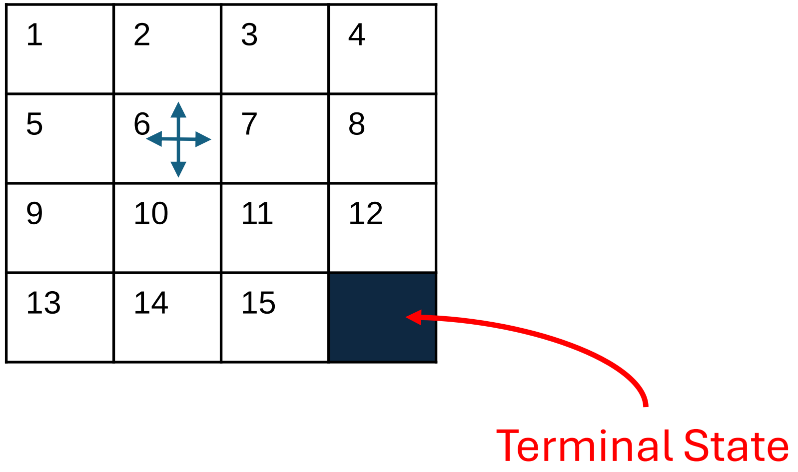
- How to formulate this maze-solving problem as an MDP?
    - States ?
    - Actions ?
    - Rewards ?
    - Transition Probabilities ?
    - Discount factor ?

How to model your problem as an MDP?

4 Actions



16 States



Rewards

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

Discount Factor

$\boldsymbol{\gamma = 1}$

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

Verify that optimal policy = shortest path

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?

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

- Hint: What discount factor will help here?