Markov Decision Processes (MDP)

Prof. Subrahmanya Swamy

user - state

articles - arms

S_t, A_t, S_{t+1}

user t

user $t+1$

Multi-Arm

- one state ✓

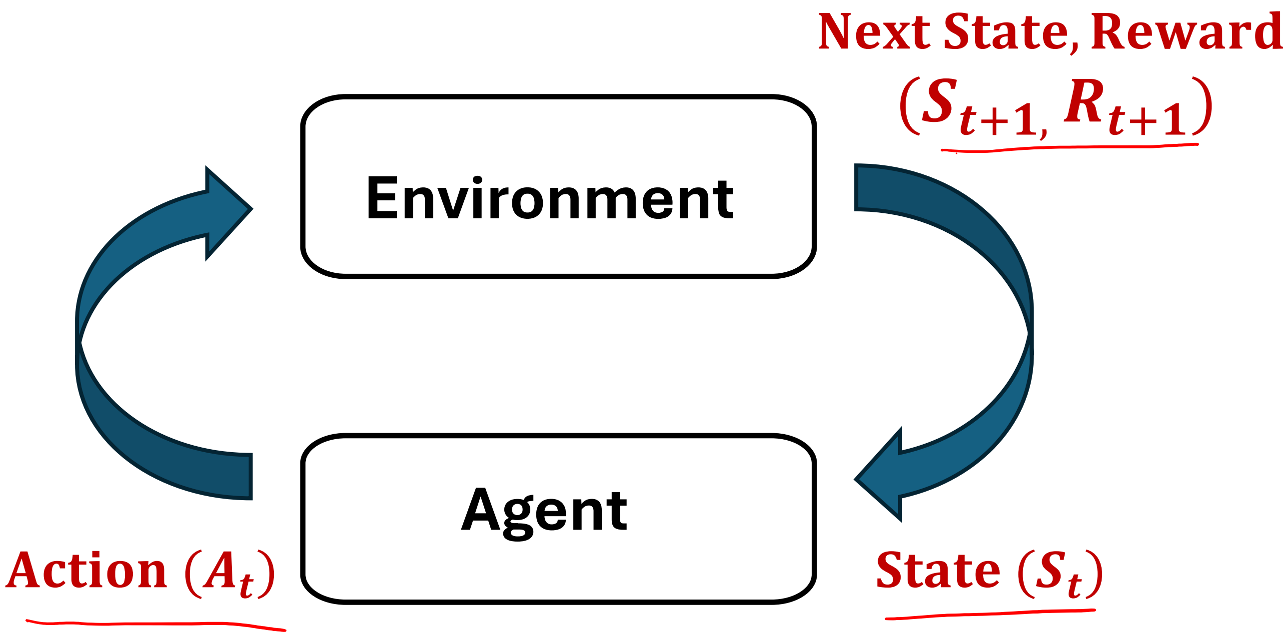
Contextual Bandits ✓

- News article

RL Framework

$S_0, A_0, R_1, S_1, A_1, R_2, \dots S_T$

$\rightarrow$ "$R_1 + R_2 + \dots R_T$"



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" $\sum_t R_t$

$\pi : S \rightarrow A$

How do we mathematically model the State transitions and Rewards?

MDP

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 → Prob (Head)

$P(X_1 = 1) = P_h$

- Knowledge of X_1 does not help in predicting X_2

X_n

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

Markov Chain

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

$1/-1$ Independent

$X_1 = 1$, $X_2 = 1$, $X_3 = -1$

- If coin lands in
  - Head: Win 1 rupee
  - Tail: Lose 1 rupee

-1 Loose

$Y_0 = 0$

$Y_2 = 1 + 1 = 2$

$Y_3 = X_3 + Y_2 = -1 + 2 = 1$

- Define Y_t = total money accumulated till time t

$Y_1, Y_2, \dots Y_t, \dots$

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

$3 + 1 = 4$

$3 - 1 = 2$

Markov Chain

$Y_1, \dots Y_t$

States = $\{0, 1, 2, \dots\}$

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

- **Markov Property:** Given the present, the future is independent of the past!

t=4

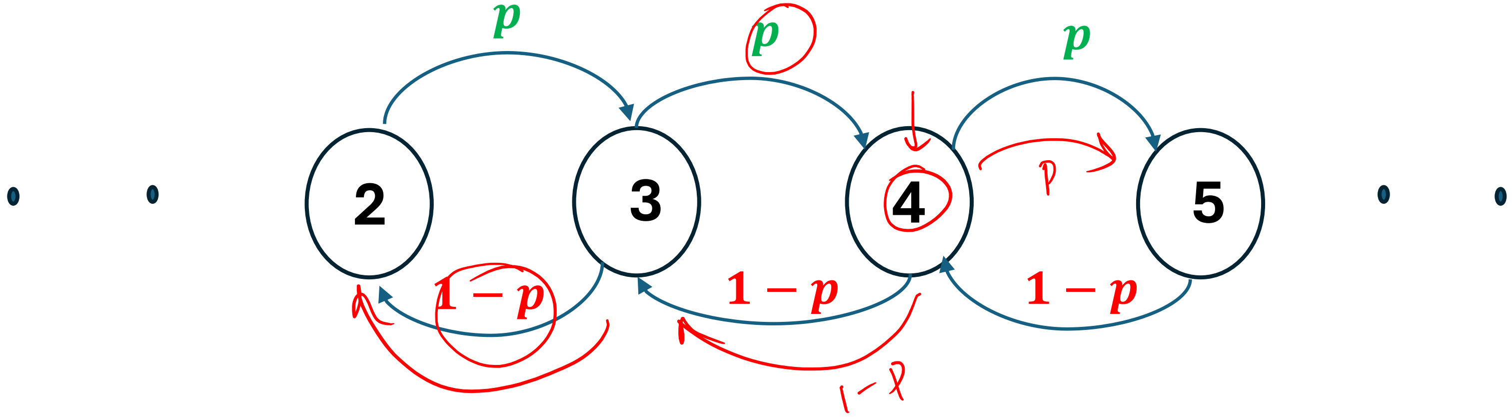
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 2, Y_3 = 3) = \frac{1}{2}$ ✓ P(Tail)
  - $\mathbb{P}(Y_5 = 1 \mid Y_4 = 2, Y_3 = 1) = \frac{1}{2}$ ✓

t=4

$Y_3 = 3$ $-1, -1$ $\to 3-2=1$

$Y_3 = 1$ $+1$ -1

$Y_4=2$ $Y_5=1$



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

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

MDP

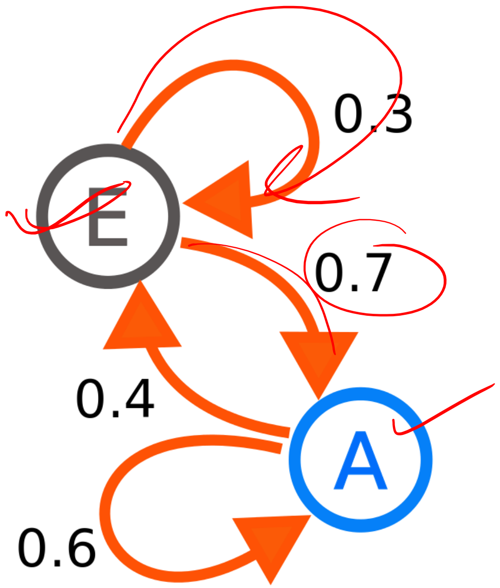
- $P_{ss'} \rightarrow Transition\ probabilty$
  - $\mathbb{P}(S_{t+1} = s' \mid S_t = s)$

$s, s' \in S$

$$\begin{array}{cc} & \begin{array}{cc} E & A \end{array} \\ \begin{array}{c} E \\ A \end{array} & \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

S_t, S_{t+1}

A_t



Markov Decision Process (MDP)

- Introduce action to convert Markov Chain into MDP

S_t =

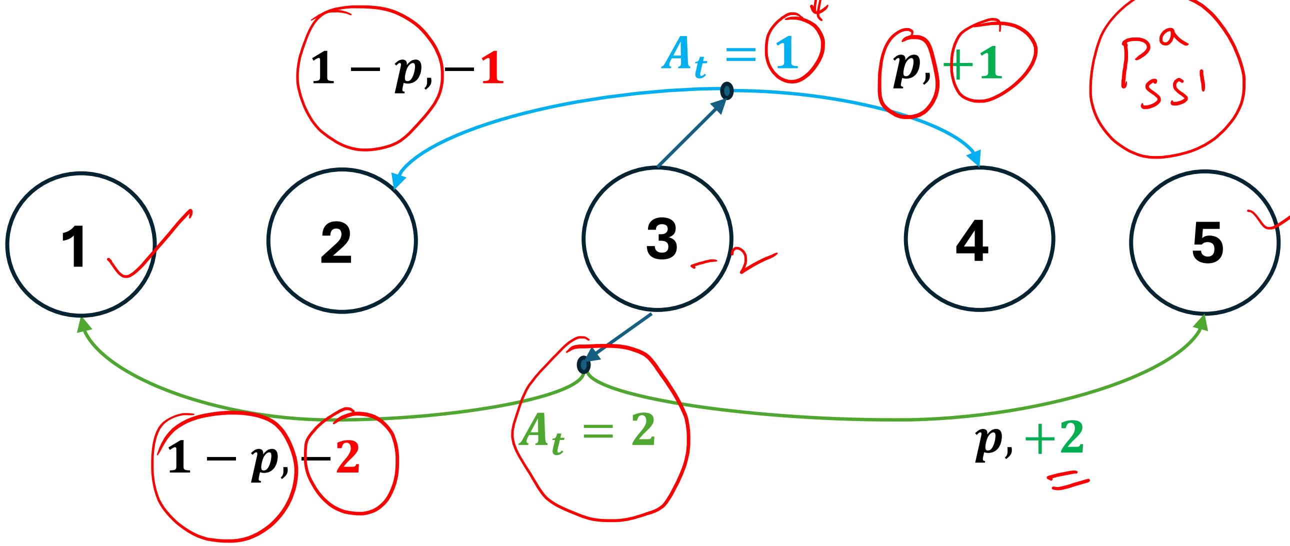
- **Actions:** How much money to bet (A_t) in the game when I have Y_t money?

$S = \{1, 2, \ldots\}$

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

$A = \{$

R



$P^a_{ss'}$

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

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

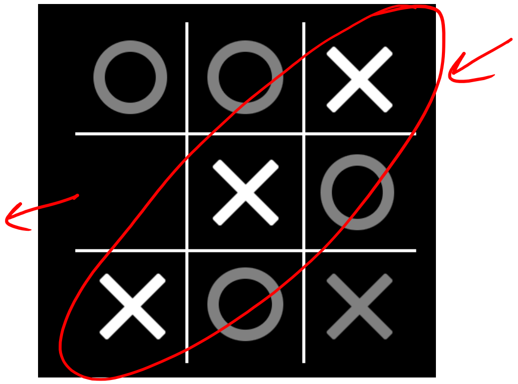
$P^2_{3,6} = 0$

Episodic and Continuing MDPs

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

→ $S_0, A_0, R_1, S_1, \ldots S_T$

Terminal state in Tic-Tac-Toe



$S_0, S_1, S_2, S_3, \ldots$

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

AMC

$S \to S \to \ldots$

Discount Factor in MDP

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

TM $\leq M = 1$

- 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 $R_i \leq M$

$\leq M$

- 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)

G.P.

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

1, 0.9, 1, $(0.9)^2$, 100

10, 0.9, $(0.99)^{100}$

$\gamma = 0.99$, $(0.001)^{100}$

$\gamma = 0.001$

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

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

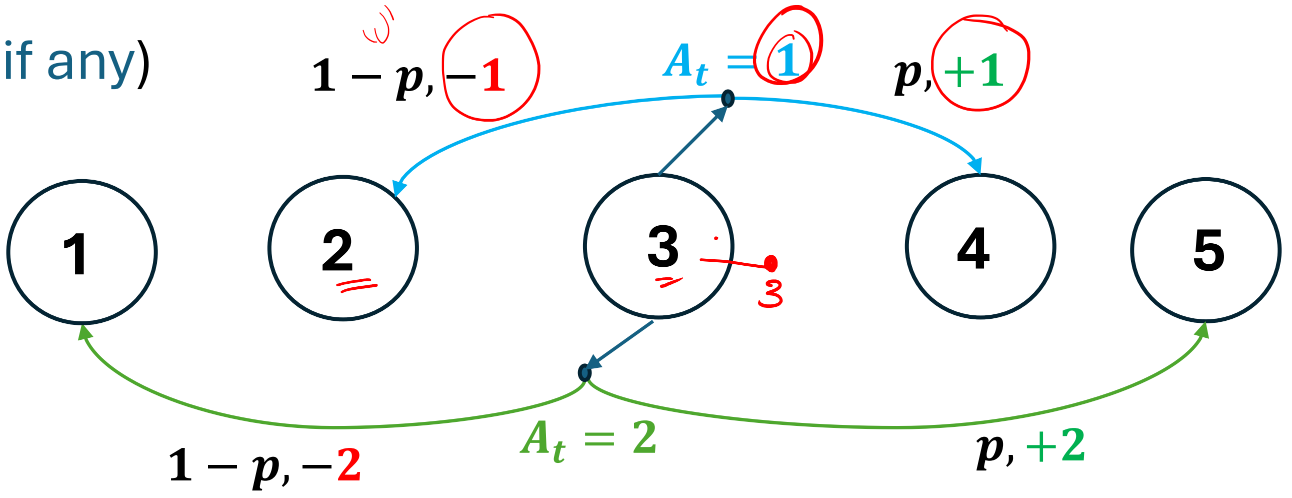
$R_3^1 = +1(P) - 1(1-P)$
$= P - 1 + P$
$= 2P - 1$

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

$E[x] = \sum_x P_x \cdot x$

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

R_s^a



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

π

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

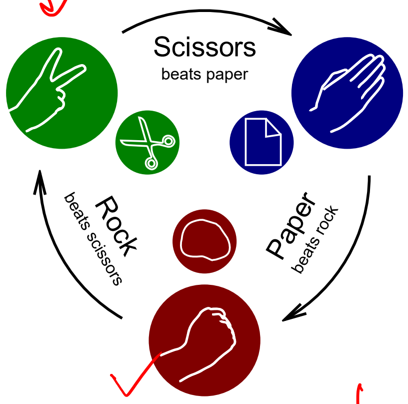
Optimal Policy

3: 1 → 1/2, 2 → 1/4, 3 → 1/4

$\pi(a|s)$

- 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

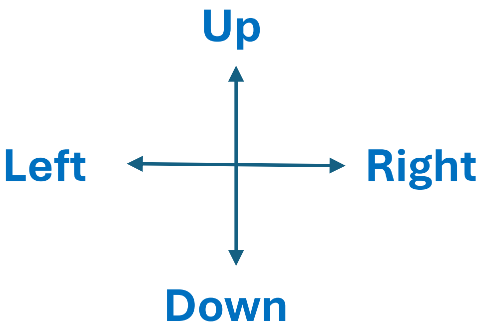
- Why stochastic policies?
  - Partially observed states
  - Exploration

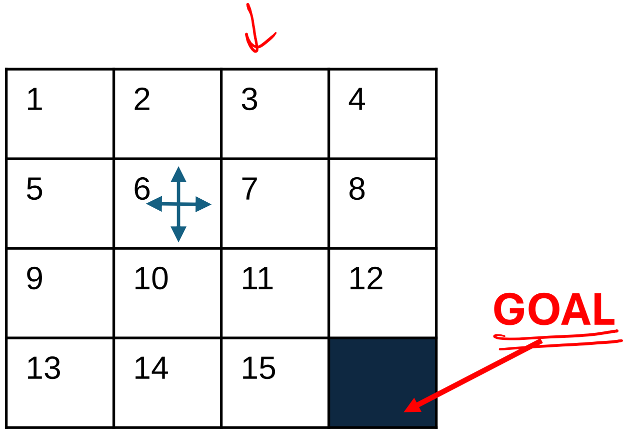


- Optimal Policy:
  - π that maximizes the expected return $\mathbb{E}_\pi[G_t \mid S_t = s]$ from any state s

How to model your problem as an MDP?

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





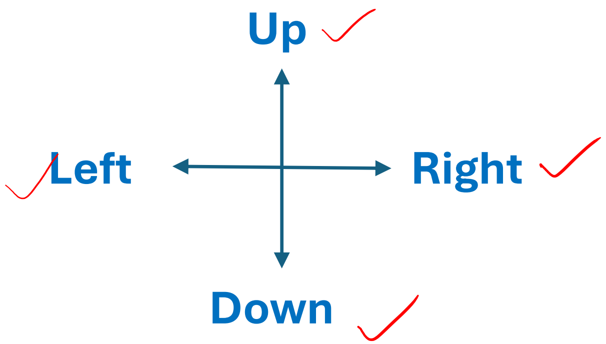
- How to formulate this maze-solving problem as an MDP?
  - States ? ✓
  - Actions ? ✓
  - Rewards ? ✓ R_s^a
  - Transition Probabilities ? ✓ $P_{ss'}^a$
  - Discount factor ? ✓ γ

How to model your problem as an MDP?

$G_t = -4$

$-1 \quad -1 \quad -1 \quad -1$

4 Actions



16 States

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

$G_t = -8$

$+1$

$= +1$

Terminal State

Rewards

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

Discount Factor

$\gamma = 1$

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

Verify that optimal policy = shortest path

$E_\pi[G_t \mid S_t = s]$

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?

$\gamma =$