Reinforcement Learning Fundamentals

Lecture 9: Markov Decision Process (MDP)

Dr Sandeep Manjanna
Assistant Professor, Plaksha University
sandeep.manjanna@plaksha.edu.in

In today's class…

- Class Presentations
  1. UCB vs. Epsilon Greedy
  2. Ridge Regression
  3. MENACE
  4. TD in Brain

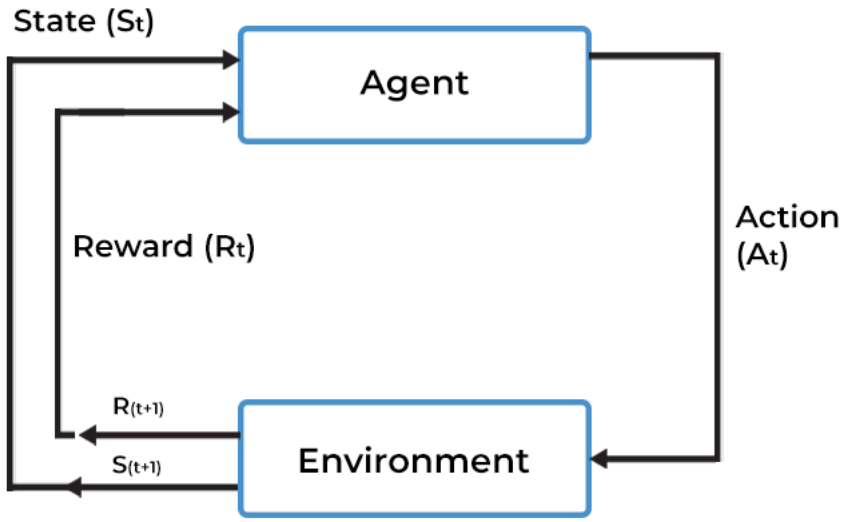
- Markov Property
- Markov Process
- Markov Reward Process
- Markov Decision Process (MDP)

Fully Observable Environments

… extent to which the agent has access to information about the current state of the environment.

- A fully observable environment is one in which the **agent has complete information about the current state of the environment.**
- The agent has direct access to all environmental features that are necessary for making decisions.
- Example?

Board games like chess or checkers.



Full observability: agent directly observes environment state

$$O_t = S_t^a = S_t^e$$

- Agent state = environment state = information state
- Formally, this is a Markov decision process (MDP)

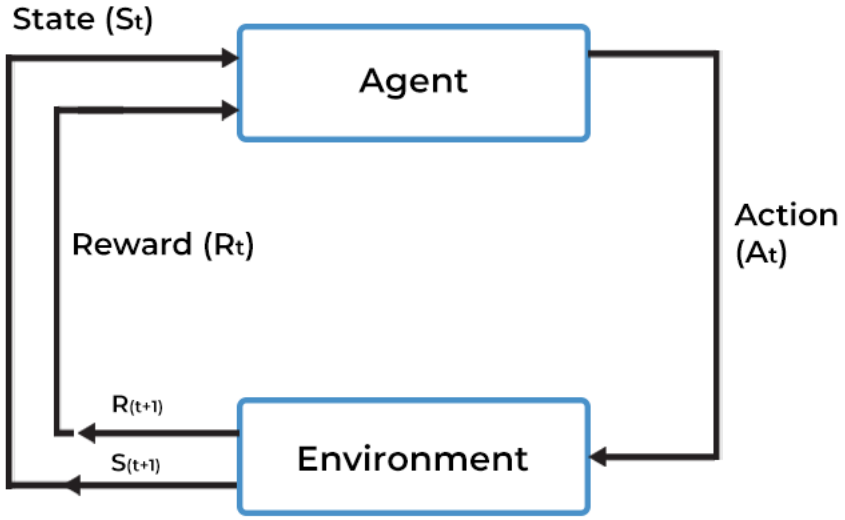
Introduction to MDP

Markov decision processes formally describe an environment for reinforcement learning

Where the environment is fully observable, i.e. The current state completely characterizes the process

- Almost all RL problems can be formalized as MDPs, e.g.
  - Optimal control primarily deals with continuous MDPs
  - Partially observable problems can be converted into MDPs
  - Bandits are MDPs with one state

Agent Environment Interface

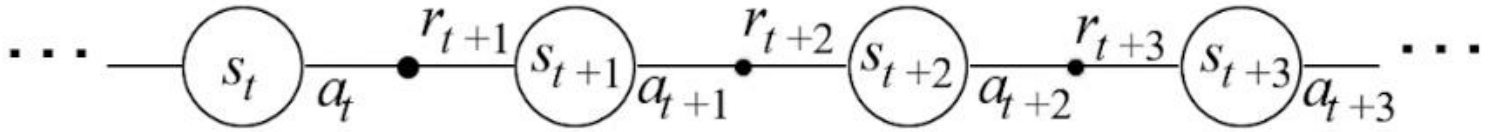


Assumption: Interaction between the agent and the environment happens at discrete time steps: t = 0, 1, 2, …

What is the maximum duration a tic-tac-toe experiment can run for?

- Agent observes state s_t at time t: $s_t \in S$
- Agent takes action at a_t time t: $a_t \in A$
- Agent gets a reward: $r_{t+1} \in \mathbb{R}$
- Agent reaches the next state: $s_{t+1} \in S$

Trajectory:



Markov Property

- “the state” at time t, means whatever information about the environment that is available to the agent at time t.
- The state can include immediate observations, highly processed observations, and structures built over time from a sequence of observations.
- Ideally, a state should summarize past observations so as to retain all essential information.
- “The future is independent of the past given the present”

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \ldots, S_t]$$

  - The state captures all relevant information from the history
  - Once the state is known, the history may be thrown away
  - i.e. The state is a sufficient statistic of the future