Reinforcement Learning Fundamentals

Lecture 11: MDP Returns and Value Function

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

In today's class…

- In-class presentations : Example MDP formulations
- Returns
- Value function

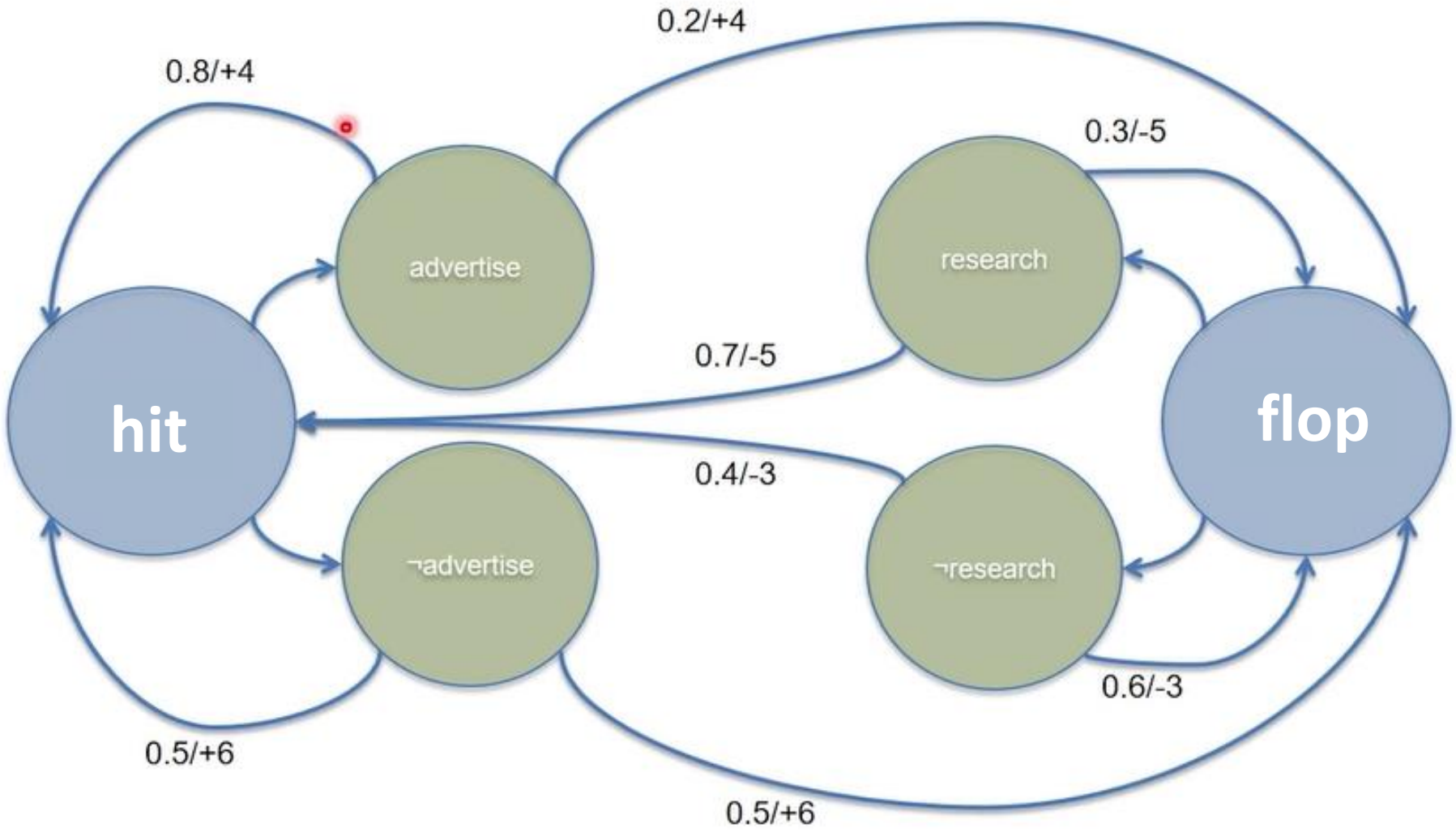
MDP Formulation

1. A web series produced by Netflix can either be a hit, or a flop. If the web series is a hit, the Netflix can advertise that show to get an immediate reward of 4 additional units. On doing so, the show continues to be a hit with a probability of 0.8. However, the show can become a flop with a probability 0.2. Netflix can alternately decide not to advertise a hit show yielding a saving of 6 units, with a 0.5 probability of the show continuing to be a hit.

   In case the show is a flop in the beginning, Netflix has an option to perform audience study to improve the viewership of the show. This investment in the study will cost the company 5 units and with a probability of 0.7, the audience study will lead to the show becoming a hit. Finally, if the company does not perform the study for the flop show, it receives an immediate cost of 3 units, with the show continuing to being a flop with a probability of 0.6.

MDP Formulation

1.



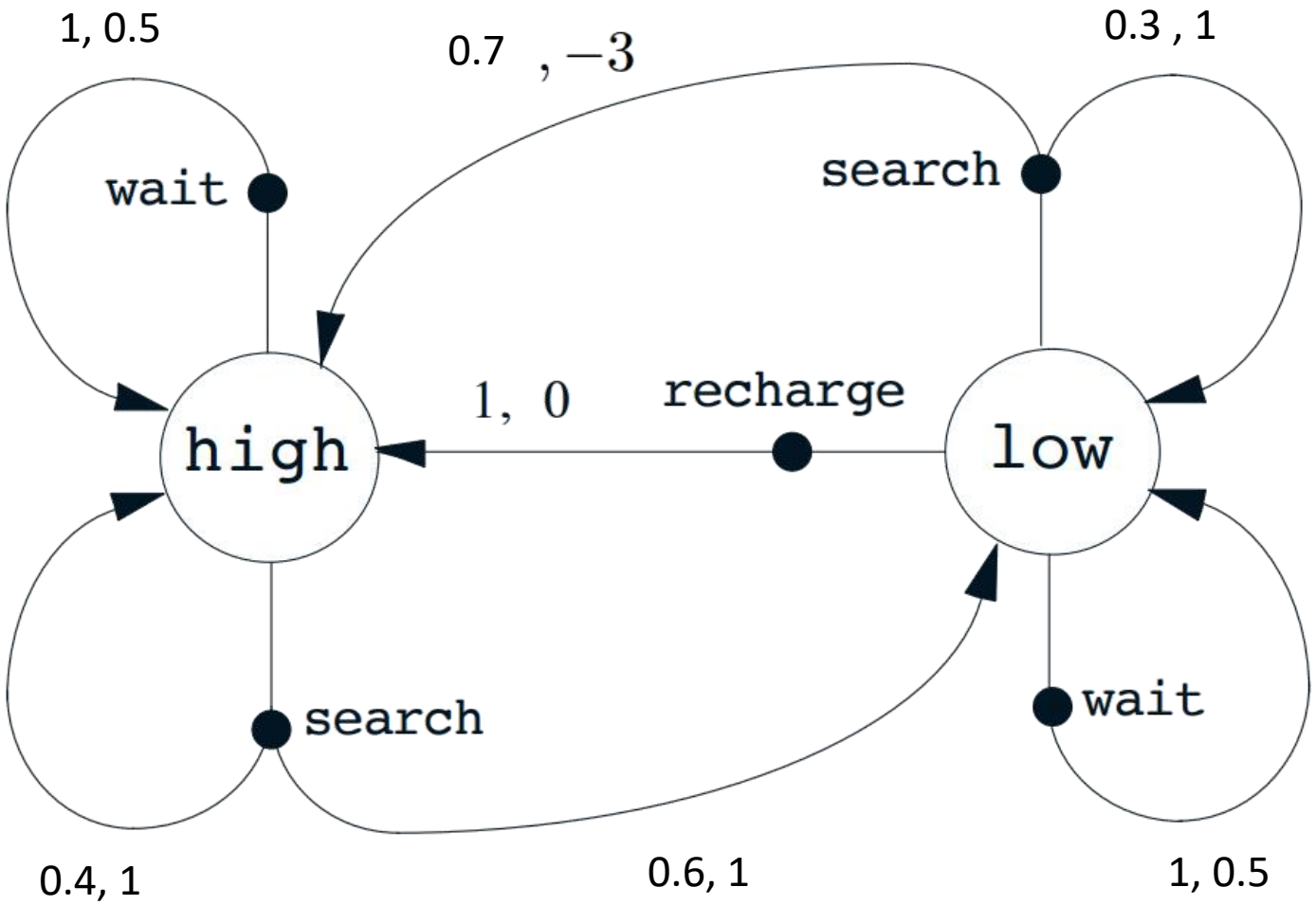
MDP Formulation

2. A mobile robot has the job of collecting empty coffee mugs in an office environment. It has sensors for detecting mugs, and an arm and gripper that can pick them up and place them in an onboard bin; it runs on a rechargeable battery. The robot's control system has components for interpreting sensory information, for navigating, and for controlling the arm and gripper. High-level decisions about how to search for mugs are made by a reinforcement learning agent based on the current charge levels (high or low) of the battery.

   The robot can wait for someone to bring the mug to it, in which case it gains an appreciation of 0.5 units. Robot with high battery level can decide to search for the mugs, receive an appreciation of 1 unit, and end up in with high battery level with a probability of 0.4. However, with a probability of 0.6, the robot's battery level goes to low. Once the robot battery is low, it can either go to the charging station to recharge its batteries, or it can decide to go searching for the mugs. But, in this case, with a probability of 0.7, the robot will drain out all the batteries and need to be carried by a human being to get the batteries charged and the robot is penalized by 3 units. And with a probability of 0.3, the robot will successfully search the mug and gain an appreciation of 1 unit.

MDP Formulation

2.



MDP Formulation

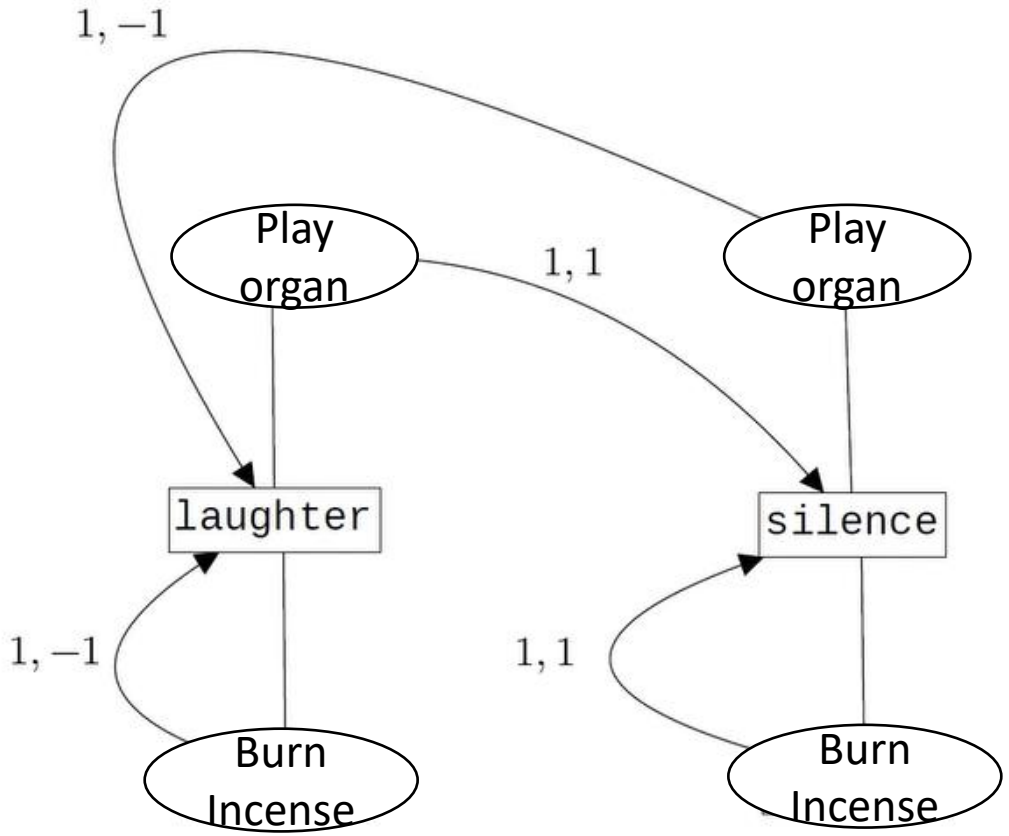
3. Dear Friend,

   Some time ago, I bought this old house, but found it to be haunted by ghostly sardonic laughter. As a result it is hardly habitable. There is hope, however, for by actual testing I have found that this haunting is subject to certain laws, obscure but infallible, and that the laughter can be affected by my playing the organ or burning incense. In each minute, the laughter occurs or not, it shows no degree. What it will do during the ensuing minute depends, in the following exact way, on what has been happening during the preceding minute: Whenever there is laughter, it will continue in the succeeding minute unless I play the organ, in which case it will stop. But continuing to play the organ does not keep the house quiet. I notice, however, that whenever I burn incense when the house is quiet and do not play the organ it remains quiet for the next minute. At this minute of writing, the laughter is going on. Please tell me what manipulations of incense and organ I should make to get that house quiet, and to keep it so.

   Sincerely,
   At Wits End

MDP Formulation

3.



More real-world examples can be found here:
https://towardsdatascience.com/real-world-applications-of-markov-decision-process-mdp-a39685546026

We want to learn a policy …

Policy at step t, π_t :

a mapping from states to action probabilities

$\pi_t(s, a) =$ probability that $a_t = a$ when $s_t = s$

- Reinforcement learning methods specify how the agent changes its policy as a result of experience.
- Roughly, the agent's goal is to get as much reward as it can over the long run.

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Returns

Suppose the sequence of rewards after step t is :

$$r_{t+1}, r_{t+2}, r_{t+3}, \cdots$$

What do we want to maximize?

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Returns

Suppose the sequence of rewards after step t is :

$$r_{t+1}, r_{t+2}, r_{t+3}, \cdots$$

What do we want to maximize?

We want to maximize the **return**, G_t, for each step t.

Episodic tasks: interaction breaks naturally into episodes, e.g., plays of a game, trips through a maze.

$$G_t = r_{t+1} + r_{t+2} + \cdots + r_T,$$

where T is a final time step at which a **terminal state** is reached, ending an episode.

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Returns

Continuing tasks: interaction does not have natural episodes.

Discounted return:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

where $\gamma, 0 \le \gamma \le 1$, is the **discount rate**.

shortsighted $0 \leftarrow \gamma \rightarrow 1$ farsighted

Will this become an Immediate RL problem?

In general,

we want to maximize the **expected return**, $E\{G_t\}$, for each step t.

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Why discount?

Most Markov reward and decision processes are discounted. Why?

- Mathematically convenient to discount rewards
- Avoids infinite returns in cyclic Markov processes
- Uncertainty about the future may not be fully represented
- If the reward is financial, immediate rewards may earn more interest than delayed rewards
- Animal/human behavior shows preference for immediate reward
- It is sometimes possible to use undiscounted Markov reward processes (i.e. $\gamma = 1$), e.g. if all sequences terminate.

Slide from David Silver's course: "Reinforcement Learning."

Value Function

- Expected future rewards starting from a state (or state-action pair) and following policy π

State - value function for policy π :

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Value Function

- Expected future rewards starting from a state (or state-action pair) and following policy π

State - value function for policy π :

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

Action - value function for policy π :

$$q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Slide from Prof. Ravindran's course: "Reinforcement Learning."

Value Function

- Expected future rewards starting from a state (or state-action pair) and following policy π

State - value function for policy π :

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

Action - value function for policy π :

$$q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$$

Slide from Prof. Ravindran's course: "Reinforcement Learning."