CS 747, Autumn 2023: Lecture 7

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2023

Markov Decision Problems

1. Alternative formulations of MDPs
2. Some applications of MDPs

Markov Decision Problems

1. Alternative formulations of MDPs
2. Some applications of MDPs

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\text{max}}, R_{\text{max}}].$$

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\text{max}}, R_{\text{max}}].$$

- You might encounter alternative definitions of R, T.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.

Reward and Transition Functions

- We had assumed
$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$
- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r|s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r|s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.
- Some authors minimise cost rather than maximise reward.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes the reward for (s, a, s') is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$, with expectation $R(s, a, s')$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r|s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.
- Some authors minimise cost rather than maximise reward.

It is relatively straightforward to handle all these variations.

Episodic Tasks

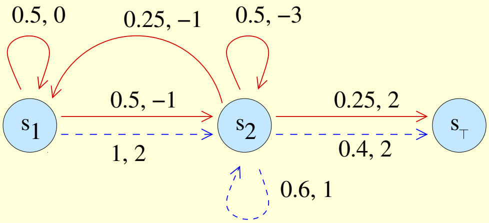
- We considered continuing tasks, in which trajectories are infinitely long.

Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.

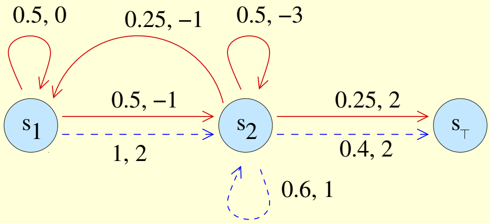
Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



Episodic Tasks

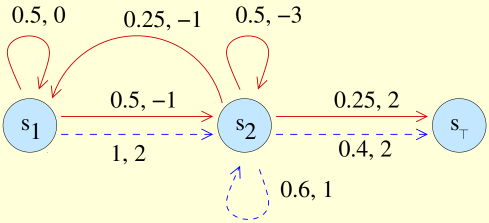
- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



- Additionally, from every non-terminal state and for every policy, there is a non-zero probability of reaching the terminal state in a finite number of steps.

Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



- Additionally, from every non-terminal state and for every policy, there is a non-zero probability of reaching the terminal state in a finite number of steps.
- Hence, trajectories or episodes almost surely terminate after a finite number of steps.

Definition of Values

- We defined $V^{\pi}(s)$ as **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.

Definition of Values

- We defined $V^{\pi}(s)$ as **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.
  There are other choices.
- **Total reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.
  Can only be used on episodic tasks.

Definition of Values

- We defined $V^\pi(s)$ as **Infinite discounted reward**:

  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.

  There are other choices.

- **Total reward**:

  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.

  Can only be used on episodic tasks.

- **Finite horizon reward**:

  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + r^1 + r^2 + \cdots + r^{H-1} | s^0 = s]$.

  Horizon $H \geq 1$ specified, rather than γ.

  Optimal policies for this setting need not be stationary.

Definition of Values

- We defined $V^{\pi}(s)$ as **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.
  There are other choices.
- **Total reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.
  Can only be used on episodic tasks.
- **Finite horizon reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \cdots + r^{H-1} | s^0 = s]$.
  Horizon $H \geq 1$ specified, rather than γ.
  Optimal policies for this setting need not be stationary.
- **Average reward** (withholding some technical details):
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[\lim_{m \to \infty} \frac{r^0 + r^1 + \cdots + r^{m-1}}{m} | s^0 = s]$.

Markov Decision Problems

1. Alternative formulations of MDPs

2. Some applications of MDPs

Controlling a Helicopter (Ng *et al.*, 2003)

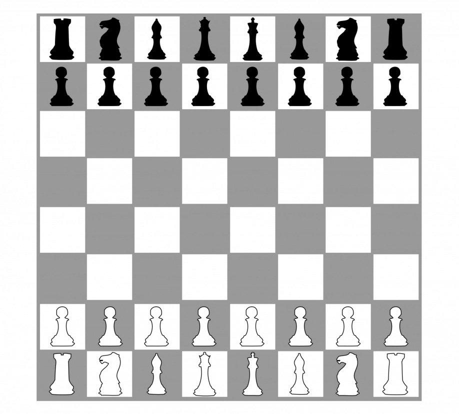
- Episodic or continuing task? What are S, A, T, R, γ?

[1]

1. https://www.publicdomainpictures.net/pictures/20000/velka/police-helicopter-8712919948643Mk.jpg.

Winning at Chess

- Episodic or continuing task? What are S, A, T, R, γ?



[1]

1. https://www.publicdomainpictures.net/pictures/80000/velka/chess-board-and-pieces.jpg.

Preventing Forest Fires (Lauer *et al.*, 2017)

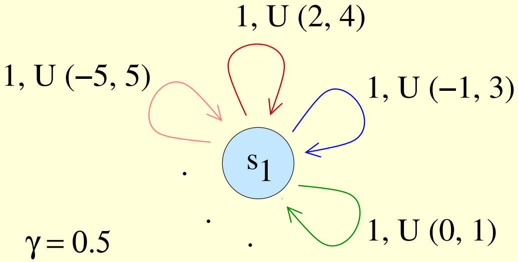
- Episodic or continuing task? What are S, A, T, R, γ?

[1]

1. `https://www.publicdomainpictures.net/pictures/270000/velka/firemen-1533752293Zsu.jpg`.

A Familiar MDP?

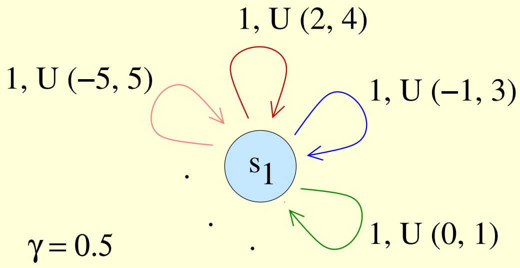
- Single state. k actions.
- For $a \in A$, treat reward of (s, a, s') as a random variable.



Annotation: "probability, reward distribution".

A Familiar MDP?

- Single state. k actions.
- For $a \in A$, treat reward of (s, a, s') as a random variable.

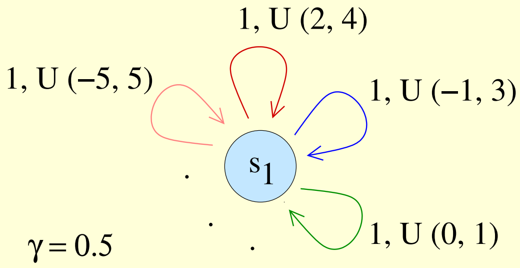


Annotation: "probability, reward distribution".

- Such an MDP is called a

A Familiar MDP?

- Single state. k actions.
- For $a \in A$, treat reward of (s, a, s') as a random variable.



Annotation: "probability, reward distribution".

- Such an MDP is called a multi-armed bandit!

Markov Decision Problems

- MDP, policy, value function
- MDP planning problem
- Policy evaluation

- Alternative formulations of MDPs
- Some applications of MDPs

- Banach's fixed point theorem
- Bellman optimality operator
- Value iteration
- Linear Programming
- Policy iteration