Markov Decision Processes

Wenhu Chen

Lecture 13

Readings: RN 17.1. PM 9.5.

Outline

Learning Goals

- Describe motivations for modeling a decision problem as a Markov decision process.
- Describe components of a fully-observable Markov decision process.
- Describe reasons for using a discounted reward function.
- Define the policy of a Markov decision process.
- Give examples of how the reward function affects the optimal policy of a Markov decision process.

Learning Goals

Introduction to Markov Decision Processes

A Grid World

Policies

The optimal policies of the grid world

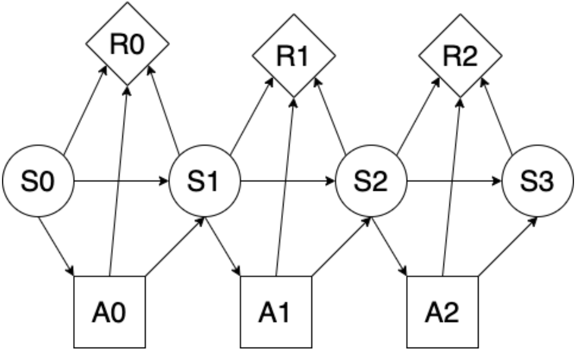
Determine the Optimal Policy Given $V^*(s)$

Revisiting Learning Goals

Modelling an Ongoing Decision Process

- Finite-stage v.s. ongoing problems
  - **Infinite horizon:** the process may go on forever.
  - **Indefinite horizon:** the agent will eventually stop, but it does not know when it will stop.
- Utility at the end vs. a sequence of rewards
  - It may not make sense to consider only the utility at the end, because the agent may never get to the end.
  - The reward incorporates the costs of actions and any rewards/punishments.

A Markov Decision Process



- S: a set of states
- A: a set of actions
- $P(s'|s,a)$ transition probabilities. A stationary model.
- $R(s,a,s')$ is the reward function.

Rewards

Define $R(s)$ as the reward received for entering state s.

- **Total reward**

$$\sum_{t=0}^{\infty} R(S_t) = R(S_0) + R(S_1) + R(S_2) + \dots$$

$\rightarrow$ If the sum is infinite, cannot compare different policies.

Rewards

Define $R(s)$ as the reward received for entering state s.

- **Total reward**

$$\sum_{t=0}^{\infty} R(S_t) = R(S_0) + R(S_1) + R(S_2) + \dots$$

$\rightarrow$ If the sum is infinite, cannot compare different policies.

- **Average reward**

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n} R(S_t) = \lim_{n\to\infty} \frac{1}{n}(R(S_0) + R(S_1) + R(S_2) + \dots)$$

$\rightarrow$ If the total reward is finite, the average reward is zero.

Rewards

- **Discounted reward**

$$\sum_{t=0}^{\infty} \gamma^t R(S_t) = R(S_0) + \gamma R(S_1) + \gamma^2 R(S_2) + \dots$$

with **discount factor** $\gamma \in [0, 1]$

Rewards

- **Discounted reward**

$$\sum_{t=0}^{\infty} \gamma^t R(S_t) = R(S_0) + \gamma R(S_1) + \gamma^2 R(S_2) + \ldots$$

  with **discount factor** $\gamma \in [0, 1]$

  $\rightarrow$ a.k.a. the **return**, $G(S_0)$

  - We prefer getting a dollar today than getting a dollar tomorrow
  - The total discounted reward is finite

    $\rightarrow$ If $\gamma \in [0, 1)$ and $R(S) \in [-R_{max}, R_{max}]$, then $\sum_{t=0}^{\infty} \gamma^t R(S_t) \leq \frac{R_{max}}{1-\gamma}$

Variations of MDP

- **Fully-observable MDP:** The state is fully observable
- **Partially observable MDP (POMDP):** combines a MDP and a hidden Markov model. The agent cannot directly observe the current state.

A 3×4 Grid World Problem

What should the robot do to maximize its rewards?

	1	2	3	4
1	Start			
2		X		-1
3				+1

- Let s_{ij} be the position in row i and column j.
- s_{11} is the initial state.
- There is a wall at s_{22}.
- s_{24} and s_{34} are goal states.
  The robot escapes the world at either goal state.

An MDP for the 3×4 Grid World

- There are four actions: *up*, *down*, *left*, and *right*.
  Every action is possible in every state.

- The transition model $P(s'|s, a)$.

  An action achieves its intended effect with probability 0.8.

  An action leads to a 90-degree left turn with probability 0.1.

  An action leads to a 90-degree right turn with probability 0.1.

  If the robot bumps into a wall, it stays in the same square.

- The reward function $R(s)$ is the reward of entering state s.

  $R(s_{24}) = -1$.

  $R(s_{34}) = 1$.

  Otherwise, $R(s) = -0.04$.

Q: Understanding the transition model

Q #1: The robot is in s_{14} and tries to move to our right, what is the probability that the robot stays in s_{14}?

(A) 0.1
(B) 0.2
(C) 0.8
(D) 0.9
(E) 1.0

	1	2	3	4
1	Start			
2		X		-1
3				+1

Q: Understanding the transition model

Q #1: The robot is in s_{14} and tries to move to our right, what is the probability that the robot stays in s_{14}?

(A) 0.1
(B) 0.2
(C) 0.8
(D) 0.9
(E) 1.0

	1	2	3	4
1	Start			
2		X		-1
3				+1

$\rightarrow$ (C) Correct answer is 0.9. If the robot moves right or up, it will stay in the current square. The total probability is $0.8 + 0.1 = 0.9$.

Q: A fixed sequence of actions

Q #2: If the environment is deterministic, an optimal solution to the grid world problem is the fixed action sequence:
down, down, right, right, right.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

Q: A fixed sequence of actions

Q #2: If the environment is deterministic, an optimal solution to the grid world problem is the fixed action sequence:
down, down, right, right, right.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

→ (A) Correct answer is True. This sequence of actions takes us to s_{34} with a minimum number of steps.

Q: A fixed sequence of actions

Q #3: Consider the action sequence *down, down, right, right, right*. This action sequence could take the robot to more than one square with positive probability.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

Q: A fixed sequence of actions

Q #3: Consider the action sequence *down, down, right, right, right*. This action sequence could take the robot to more than one square with positive probability.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

→ (A) Correct answer is True. For example, it could take us to s_{12} if we end up going right, down, down, down, and down.

A policy must specify what to do in every state.

Policies

A **policy** specifies what the agent should do as a function of the current state.

A policy is

- *non-stationary* if it is a function of the state and the time.
- *stationary* if it is a function of the state.

The optimal policies of the grid world

→ Before I show you the algorithm to solve an MDP (next lecture), let's look at how the optimal policy of the grid world is influenced by the reward function $R(s)$.

The optimal policy of the grid world changes based on $R(s)$ for any non-goal state s. It shows a careful balancing of risk and reward.

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

Reward – want to reach the $+1$ state ASAP. There are 2 paths.

Risk – do not want to fall into the -1 state. The longer we explore, we more negative penalty we accumulate.

The optimal policy when life is ...

Let's determine optimal policies for these situations, where the reward function for entering nonterminal states is:

- $R(s) < -1.6284$
- $-0.4278 < R(s) < -0.0850$
- $R(s) = -0.04$
- $-0.0221 < R(s) \leq 0$
- $0 < R(s)$

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

The optimal policy when life is quite unpleasant

When $-0.4278 < R(s) < -0.0850$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				+1

$\rightarrow$ Optimal policy:

	1	2	3	4
1	$\downarrow$	$\rightarrow$	$\downarrow$	$\leftarrow$
2	$\downarrow$	X	$\downarrow$	-1
3	$\rightarrow$	$\rightarrow$	$\rightarrow$	+1

Life is quite unpleasant. The agent takes the shortest route to the $+1$ state and is willing to risk falling into the -1 state by accident. The agent takes the shortcut from s_{13}.

The optimal policy when life is painful

When $R(s) < -1.6284$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

$\rightarrow$ Optimal policy:

	1	2	3	4
1	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\downarrow$
2	$\downarrow$	X	$\rightarrow$	-1
3	$\rightarrow$	$\rightarrow$	$\rightarrow$	$+1$

Life is so painful that the agent heads straight for the nearest exit, even if the exit is worth -1.

The optimal policy when life is unpleasant

When $R(s) = -0.04$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				+1

$\rightarrow$

	1	2	3	4
1	$\downarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
2	$\downarrow$	X	$\downarrow$	-1
3	$\rightarrow$	$\rightarrow$	$\rightarrow$	+1

The optimal policy for s_{13} is conservative. We prefer to take the long way around to avoid reaching the -1 state by accident.

The optimal policy when life is only slightly dreary

When $-0.0221 < R(s) \leq 0$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				$+1$

$\rightarrow$

	1	2	3	4
1	$\downarrow$	$\leftarrow$	$\leftarrow$	$\uparrow$
2	$\downarrow$	X	$\leftarrow$	-1
3	$\rightarrow$	$\rightarrow$	$\rightarrow$	$+1$

Life is only slightly dreary. The optimal policy takes no risk.

In s_{14} and s_{23}, the agent heads directly away from the -1 state to avoid falling into the -1 state by accident even though this means banging its head against the wall quite a few times.

The optimal policy when life is GOOD =D

When $R(s) > 0$,
what does the optimal policy look like?

	1	2	3	4
1	Start			
2		X		-1
3				+1

$\rightarrow$

	1	2	3	4
1	$\uparrow\downarrow\leftarrow\rightarrow$	$\uparrow\downarrow\leftarrow\rightarrow$	$\uparrow\downarrow\leftarrow\rightarrow$	$\uparrow$
2	$\uparrow\downarrow\leftarrow\rightarrow$	X	$\leftarrow$	-1
3	$\uparrow\downarrow\leftarrow\rightarrow$	$\uparrow\downarrow\leftarrow\rightarrow$	$\leftarrow$	+1

Life is so pleasant and the agent avoids both goal states.

Q: A stochastic environment

Q #4: True or False: The solution to this problem should be a fixed sequence of actions. For example, a fixed sequence of actions is *down, down, right, right, right*.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

Q: A stochastic environment

Q #4: True or False: The solution to this problem should be a fixed sequence of actions. For example, a fixed sequence of actions is *down, down, right, right, right*.

(A) True
(B) False
(C) I don't know

	1	2	3	4
1	Start			
2		X		-1
3				+1

→ (B) Correct answer is False. Since actions may not have their intended effects, a fixed sequence of actions is not guaranteed to reach a desirable state. We need to constantly revise our plan based on the current state we are in.

The Expected Utility of a Policy

$V^{\pi}(s)$: expected utility of entering state s and following the policy π thereafter.

$V^{*}(s)$: expected utility of entering state s and following the optimal policy π^{*} thereafter.

The Values of $V^*(s)$

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	$+1$

Figure: $V^*(s)$ for $\gamma = 1$ and $R(s) = -0.04, \forall s \neq s_{24}, s \neq s_{34}$.

Calculate the Optimal Policy Given $V^*(s)$

Calculate my expected utility if I am in state s and take action a.

$$Q^*(s, a) = R(s) + \sum_{s'} P(s'|s, a) V^*(s') \tag{1}$$

In state s, choose an action that maximizes my expected utility.

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{2}$$

Q: Determine optimal action given $V^*(s)$

Q #5: What is the optimal action for state s_{13}?

(A) Up (B) Down (C) Left (D) Right

$$Q^*(s,a) = R(s) + \sum_{s'} P(s'|s,a)V^*(s')$$
$$\pi^*(s) = \arg\max_a Q^*(s,a).$$

The values of $V^*(s)$ are given below.

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	+1

Q: Determine optimal action given $V^*(s)$

$$Q^*(s,a) = R(s) + \sum_{s'} P(s'|s,a)V^*(s')$$
$$\pi^*(s) = \arg\max_a Q^*(s,a).$$

The values of $V^*(s)$ are given below.

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	+1

Q: Determine optimal action given $V^*(s)$

$$Q^*(s,a) = R(s) + \sum_{s'} P(s'|s,a)V^*(s')$$
$$\pi^*(s) = \arg\max_a Q^*(s,a).$$

The values of $V^*(s)$ are given below.

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	+1

Left: -0.04 + 0.8 * 0.655 + 0.1 * 0.611 + 0.1 * 0.660 = 0.6111
Right: -0.04 + 0.8 * 0.388 + 0.1 * 0.611 + 0.1 * 0.660 = 0.3975
Down: -0.04 + 0.8 * 0.660 + 0.1 * 0.388 + 0.1 * 0.655 = 0.5923
Up: -0.04 + 0.8 * 0.611 + 0.1 * 0.655 + 0.1 * 0.388 = 0.5531

Q: Determine optimal action given $V^*(s)$

Q #5: What is the optimal action for state s_{13}?

(A) Up (B) Down (C) Left (D) Right

$$Q^*(s,a) = R(s) + \sum_{s'} P(s'|s,a)V^*(s')$$

$$\pi^*(s) = \arg\max_a Q^*(s,a).$$

The values of $V^*(s)$ are given below.

	1	2	3	4
1	0.705	0.655	0.611	0.388
2	0.762	X	0.660	-1
3	0.812	0.868	0.918	+1

$\rightarrow$ Answer is (C), going left

Revisiting Learning Goals

- Describe motivations for modeling a decision problem as a Markov decision process.
- Describe components of a fully-observable Markov decision process.
- Describe reasons for using a discounted reward function.
- Define the policy of a Markov decision process.
- Give examples of how the reward function affects the optimal policy of a Markov decision process.