MITx 6.86x

Machine Learning with Python-From Linear Models to Deep Learning

Course Progress Dates Discussion Resources

Course / Unit 5. Reinforcement Learning (2 we... / Lecture 17. Reinforcement L

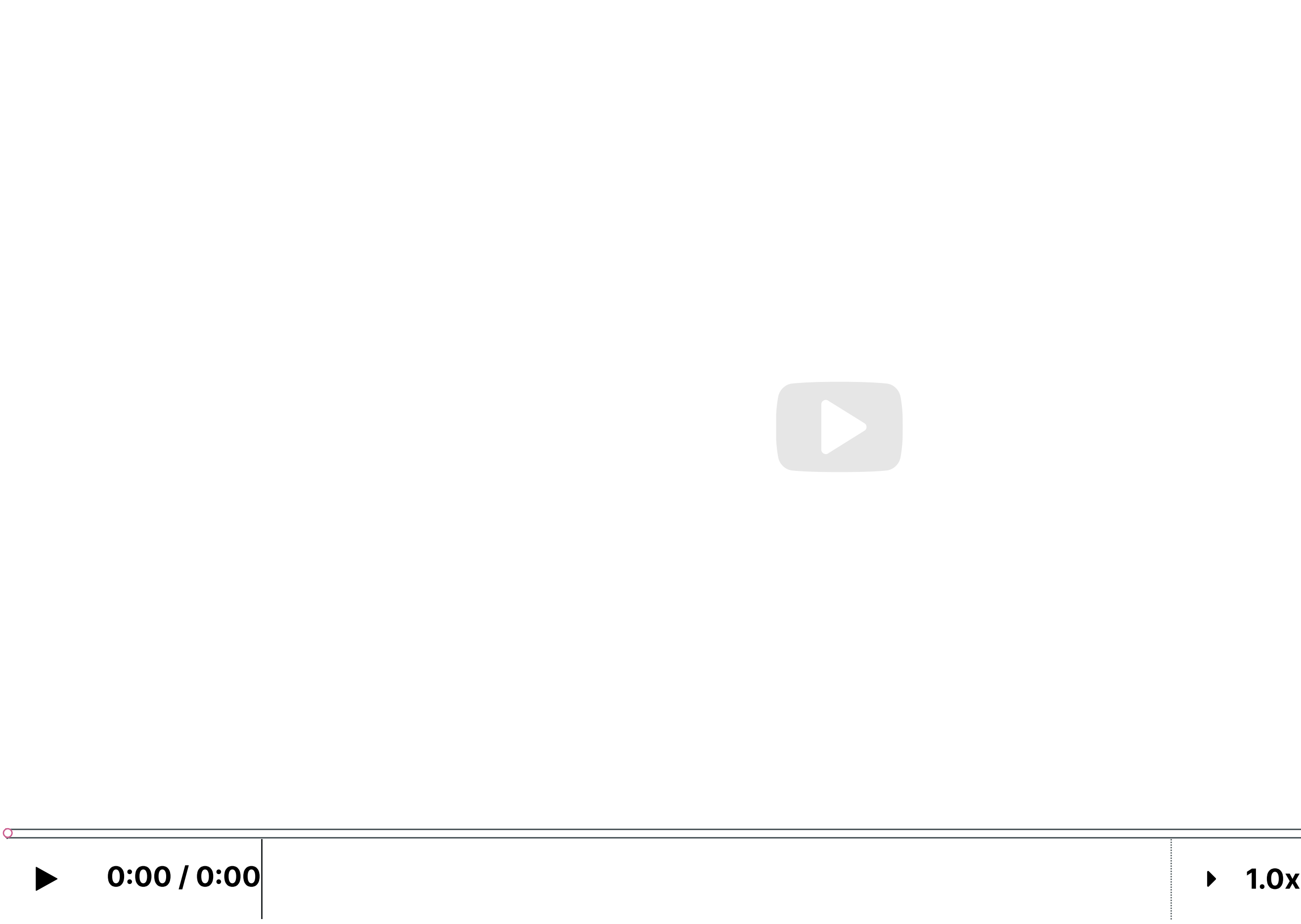
‹ Previous ✓ ✓ ✓ ✓ ✓ ✓

5. Policy and Value Functions

Bookmark this page

Exercises due May 3, 2023 08:59 -03 Completed

Policy and Value Functions



Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Definition of Optimal Policy

1/1 point (graded)

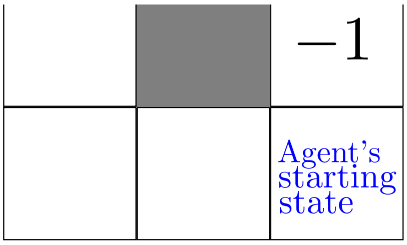
Given an MDP, and a utility function $U\left[s_0, s_1, \ldots, s_n\right]$, , our goal is to find an optimal maximizes the expectation of the utility. Here, a **policy** is a function $\pi : S \to A$ that as any state s. We denote the optimal policy by π^*.

Which of the following option is correct about the optimal policy function?

○ The optimal policy function would only depend on the state and action space bu reward structure.

● The optimal policy assigns an action at every state that maximizes the expected

○ For any given state, the optimal policy function should always take an action tha



Recall the MDP example in the lecture. An AI agent navigates in the 3×3 grid depicted a
square is not accessible (and hence is greyed out).

The MDP is defined as follows. As before, every state is defined by the current positi
grid. The actions are the 4 directions “up", “down",“left", “right".

Now, The transition probabilities from state via action to state is given by

Reward structure:

As before, the agent receives a reward of for arriving at the top right cell, and a rew
the cell immediately below it. It does not receive any non-zero reward at the other cells
following figure.

However, this time, the agent also receives a reward (or penalty) of for every actio
any action that leads the agent into the or cells.

Transition Probabilities:

For simplicity, assume that all the transitions are deterministic. That is, given any state
actions are deterministic: The next state reached is completely predictable.

For intance, taking the action “left" from the bottom right cell will always take the agent
to its left. Any action pointing off the grid would lead the agent to remain in its current c

Initial State:

Also, assume that the agent always starts off from the bottom right corner of the grid.
action until it reaches the top right corner, at which point it stops and does not act anyr

Optimal policy - Numerical Example

2/2 points (graded)

Recall that in this setup, the agent receives a reward (or penalty) of for every actio
of the and when it reached the corresponding cells. Since the agent always sta
the outcome of each action is deterministic, the discounted reward depends only on th
can be written as:

Maximum discounted reward: -15.5 ✓

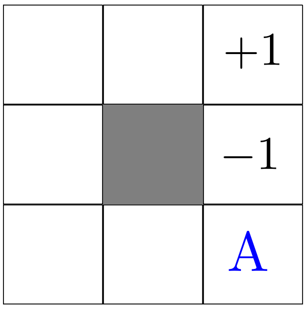
Submit You have used 1 of 3 attempts

Value Function

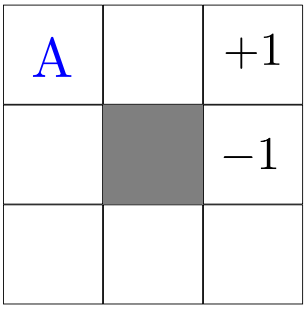
0/1 point (graded)

As above, we are working with the grid example with reward at the top right cell below it. The agent also gets a reward of for every action that it takes. The ac deterministic. The agent continues to act until it reaches the cell, when it stops.

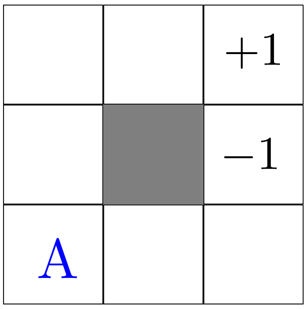
The following figures show states in which the letter “A" marks the current lo



s_1



s_2



s_3

A **value function** of a given state is the expected reward (i.e the expectation o the agent acts optimally starting at state . In the given MDP, since the action outcome expected reward simply equals the utility function.

Which of the following should hold true for a good value function under the rewa MDP?

Note: You may want to watch the video on the next page before submitting this questio

●

○

○

✗

< Previous Next >

I believe that the equation in the second part in not mathematically correct. Why does the left side depends

? Value Function - is gamma assumed to be 1 in the last question?

* didn't see gamma in the answer for that question

? [STAFF] - Optimal policy - Numerical Example - last step

This is a critical for my understanding. Why we do not count last step as gamma^2*R(s_2,a_3), a_3=0 ????? If

💬 Better example needed

The study example needs to be refined or even better replaced with a more intuitive one. It is not a good one

? Formal definition of R for the numerical problem

How does one define R? The description doesn't really define it formally. Specifically, does the R(s(n),a(n+1))

💬 s_0

💬 awful notation and word choice

? Optimal policy - ambiguity in question

The question clearly states: "For the cases (gamma=0) and (gamma=0.5), what is the maximum discounted r

? Confusion about what step the +1/-1 rewards apply

Based on the discounted rewards function, I am not clear at which term the destination rewards of +1 and -1

? Answer for gamma = 0.5 does not seem correct

The answer for gamma = 0.5 does not seem correct, as it didn't factor in R(s_2, a_3).

Legal

Terms of Service & Honor Code

Privacy Policy

Accessibility Policy

Trademark Policy

Sitemap

Cookie Policy

Do Not Sell My Personal Information

Connect

Blog

Contact Us

Help Center

Security

Media Kit