Audit Access Expires May 11, 2020
You lose all access to this course, including your progress, on May 11, 2020.

1. Value Iteration for Markov Decision Process

Consider the following problem through the lens of a Markov Decision Process (MDP), and answer questions 1 - 3 accordingly.

Damilola is a soccer player for the ML United under-15s who is debating whether to sign for the NLP Albion youth team or the Computer Vision Wanderers youth team. After three years, signing for NLP Albion has two possibilities: He will still be in the youth team, earning 10,000 (60% chance), or he will make the senior team and earn 70,000 (40% chance). Lastly, he is assured of making the Computer Vision Wanderers senior team after three years, with a salary of 37,000.

Q1

1 point possible (graded)

Given that Damilola only cares about having the highest expected salary after three years, $V^*(\text{ML United under-15s})$ is achieved through the action of signing for Computer Vision Wanderers.

- True ✔
- False

Solution:

$37,000 > 0.6 * 10,000 + 0.4 * 70000 = 34,000.$

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Q2

1 point possible (graded)

Let us now assume that Damilola cares about the utility derived from the salary as opposed to the salary S itself. And his utility function, which baffles economists, is given by Utility, $U = \Psi S^2 + \zeta$ where $\Psi, \zeta > 0$, and Ψ & ζ are constants. In this scenario, the optimal policy π^* (ML United under-15s) would be to sign for NLP Albion.

- True ✔
- False

Solution:

Since Ψ and ζ are constants, we only need to compare the S^2 terms:
$0.6 * (10,000^2) + 0.4 * (70,000^2) = 2.020 \times 10^9 > 37,000^2 = 1.369 \times 10^9$
.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Q3

1 point possible (graded)
There are 3 unique states defined in total in this setting.

- True
- False ✔

Solution:

There are a total of 5 states: ML United under-15s, NLP Albion youth team, NLP Albion senior team, Computer Vision Wanderers youth team, and Computer Vision Wanderers senior team.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Convergence of the Value Iteration Algorithm

1 point possible (graded)
For an Markov Decision Process (MDP) with a single state and a single action, we know the following hold:

$$\begin{aligned} V_{i+1} &= R + \gamma V_i \\ V^* &= R + \gamma V^* \end{aligned}$$

Working with these equations, we can conclude that after each iteration, the difference between the estimate and the optimal value of V is decreased by how much? Answer by finding C in the equation below:

$$(V_{i+1} - V^*) = C\,(V_i - V^*)$$

.

(Enter your answer in terms of γ.)

$C =$ **Answer:** gamma

STANDARD NOTATION

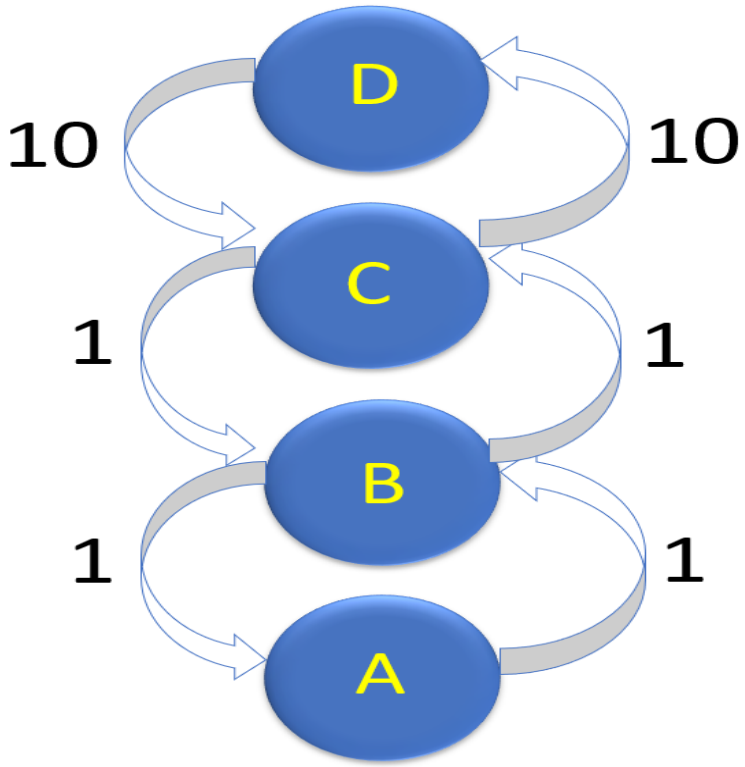
Solution:

Subtracting equation (2) from equation (1) gives: $(V_{i+1} - V^*) = \gamma\,(V_i - V^*)$, which shows us that after each iteration the difference between our estimate and the optimal value decreases by γ.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Consider the following Markov Decision Process (MDP):



MDP with 4 states (rewards for each action are indicated on the arrow)

There are 4 states A, B, C, and D. We can move up or down from states B and C, but only up for A and only down for D. Note that the discount factor $\gamma = 0.75$, and that this MDP is deterministic i.e. if you choose action UP, you are guaranteed to move UP, and likewise for action DOWN.

---

a

4 points possible (graded)

What are the optimal policies for each state?

$\pi^*(A) =$

- (●) UP ✔
- ( ) DOWN

$\pi^*(B) =$

- (•) UP ✔
- ( ) DOWN

$\pi^*(C) =$

- (•) UP ✔
- ( ) DOWN

$\pi^*(D) =$

- ( ) UP
- (•) DOWN ✔

Solution:

Explanation: For state A, it is evident that the optimal policy is to move UP, as you cannot move down. And for state D, the optimal policy is DOWN as we cannot move up. For states B and C, we can move both UP and DOWN, but moving UP is the optimal choice for each of these states. This is because the rewards associated with state D at the top are higher than the rewards associated with state A at the bottom.

Submit You have used 0 of 1 attempt

ℹ Answers are displayed within the problem

b

3 points possible (graded)
If we initialize the value function with 0, enter the value of state B after:

one value iteration, V_{B1}^* **Answer:** 1

two value iterations, V_{B2}^* **Answer:** 8.5

infinite value iterations, V_B^* **Answer:** 31

Solution:

The reward for moving up from B to C is 1.
With 2 iterations, $1 + 0.75 * 10 = 8.5$.
With infinite number of iterations, $1 + (\sum_{i=1}^{\infty} 0.75^i) * 10 = 31$.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

C

1 point possible (graded)
Select all that are true

- [ ] In an MDP, the optimal policy for a given state s is unique
- [ ] The value iteration algorithm is solved recursively ✔
- [ ] For a given MDP, the value function $V^*(s)$ of each state is known a priori
- [ ] $V^*(s) = \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma V^*(s')]$
- [ ] $Q^*(s,a) = \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma V^*(s')]$ ✔

Solution:

There can be multiple optimal policies for a given state.
The value iteration algorithm recursively estimates $V_k^*(s)$.
The $V(s)$ are not known a priori - they are found by the value iteration algorithm.
$V^*(s) = \max_a \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma V^*(s')]$.
$Q^*(s,a) = \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma V^*(s')]$

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Homework 6 /
1. Value Iteration for Markov Decision Process

Add a Post

Show all posts by recent activity

[staff] Q3 Ambiguity - need clarification & counter reset 8
Hi Staff, With all due respect, I think it will be beneficial either to remove this problem for goo...

☑	Solved the MDP problem correctly with Excel, but looking for a Vector solution I couldn't figure out how to correctly set this up using Vector operations. I tried using the sam...	5
?	having some trouble with value iteration Convergence of the Value Iteration Algorithm part b I managed to get V_B1 and V_B2 but don...	5
?	Q3. Clarification required I am hoping that I can get clarification for question Q3 without giving away too much. Should ...	8
?	[Staff] Q. b - notation	2
☑	Confusion Formula given is - > Vi+1 =R+γVi But then there is another formula given where Vi+1 is depen...	6
☑	edX issue for Q4 - For an MDP with 1 state and 1 action I was about to submit my answer but it failed. When I refreshed the page, I got: Could not for...	3
☑	@staff: Question on Markov Decision Process (MDP) part B Used the code similar to the exercise to solve the vb1* ,vb2*, vb*. Got 1 and 2 correct but vb ...	5
☑	For b, the probability of B moving up and down is both 0.5? as title	3
?	[@STAFF] : request to reset attempts request to reset attempts please	2
💬	Damilola - a beautiful Nigerian name That's a beautiful Nigerian name. I was really surprised but glad to see this :).	2
💬	Tips for vectorization Define Transition matrix. (4 * 4) with values 1 if its possible to go from state x to state y and 0...	5
☑	[staff]: The problem C - need clarification	12