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

1/1 point (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 1 of 1 attempt

Answers are displayed within the problem

Q2

1/1 point (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 $\pi^*(\text{ML United under-15s})$ 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 1 of 1 attempt

Answers are displayed within the problem

Q3

1/1 point (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.

You have used 1 of 1 attempt

Submit

Answers are displayed within the problem

Convergence of the Value Iteration Algorithm

1.0/1 point (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 decreases by a factor of ? (Enter your answer in terms of γ)

gamma ✔ **Answer:** gamma

STANDARD NOTATION

Solution:

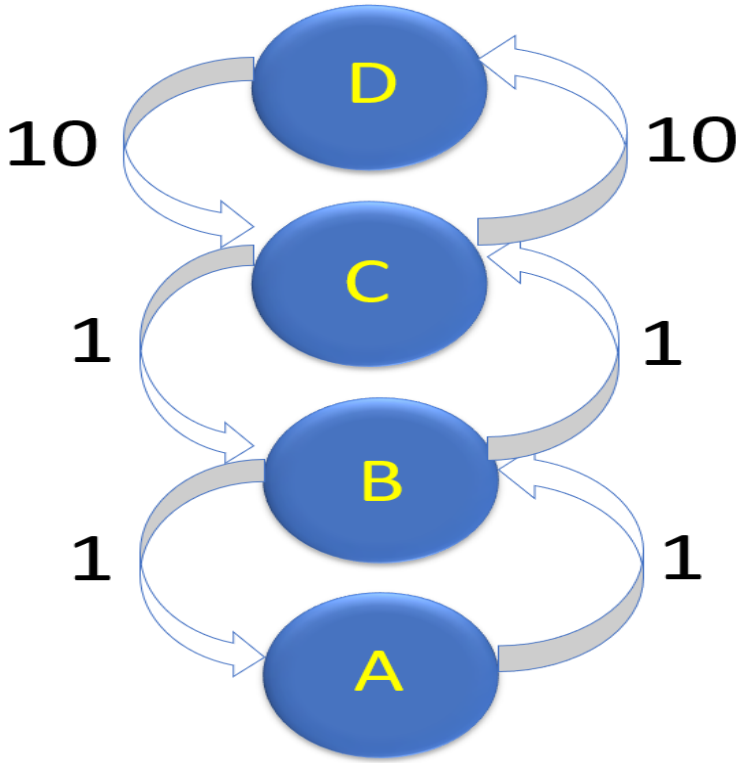
Subtracting equation (2) from equation (1) gives: $(V_{i+1} - V^*) = \gamma (V_i - V^*)$, which shows us that after each iteration the difference between our estimate and the optimal value decreases by γ.

You have used 1 of 3 attempts

Submit

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 y= 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

4/4 points (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 1 of 1 attempt

Answers are displayed within the problem

b

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

one value iteration, V_{B1}^* 1.0 ✔ **Answer:** 1

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

infinite value iterations, V_B^* 31 ✔ **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 2 of 3 attempts

Answers are displayed within the problem

C

1/1 point (graded)
Select all that are true

- [ ] In an MDP, the set of optimal policies for a given state s is a singleton
- [x] The value iteration algorithm is solved recursively ✔

- [ ] For a given MDP, the value of each state is known a priori
- [ ] $V^*(s) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$
- [x] $Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$ ✔

✔

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.
$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$.
$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$

Submit You have used 1 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 ▼

☑	3b: Doesn't our current formulation diverge?	37
💬	[STAFF] "The value iteration algorithm is solved recursively" An algorithm may be a recursive method to solve an equation or other problem. An algorithm is not itself solved.	5
?	MDP question b, infinite iterations.	3
💬	@STAFF last minute chance please! In Q3-c, I clicked on submit after the first try and one of my chances was gone. I appreciate if you give me another chance, please.	2
💬	Baffled economist here So I suppose I should be baffled because our Damilola has increasing marginal returns to a higher salary? Must have a fierce desire to become ri...	5
💬	was looking forward to RL part but IMHO it's by far the worst one too many vague definitions very shallow coverage tasks are ill defined and can be interpreted in multiple ways hugely disappointed can anybody ...	3
💬	I believe the RL part should be reformed for the next semesters While I have enjoyed the rest of the course, this unit has been far below my expectations. I think videos were too brief and superficial and questi...	2
💬	@Staff May I have one more try? in Q3-c, my first and second tries were the same because I clicked on submit by mistake. Could you give me another try?	1
?	[staff] I clicked in submit button unintentionally I clicked in submit button of last question unintentionally. You can see that the answer I gave in my last and unintentional try is the same I had gi...	2
💬	Q1: Does highest possible salary here refer to highest expected salary? Q1: Just wanted to confirm whether "highest possible salary" here refers to highest expected salary.(we are supposed to consider the risk or T al...	2 new

★ Following

Q4 - "decrease by a factor" 5

Community TA

Q2 16

Cannot see my progress/grade chart! 3

Hello team! I am going to the progress bar but the grade chart that used to appear is no longer there. What has happened?