2. Q-Value Iteration

Consider an Markov Decision Process with 6 states $s \in \{0, 1, 2, 3, 4, 5\}$ and 2 actions $a \in \{C, M\}$, defined by the following transition probability functions

For states 1, 2, and 3:

$$T(s, M, s-1) = 1$$

$$T(s, C, s+2) = 0.7$$

$$T(s, C, s) = 0.3$$

For state 0:

$$T(s, M, s) = 1$$

$$T(s, C, s) = 1$$

For states 4 and 5:

$$T(s, M, s-1) = 1$$

$$T(s, C, s) = 1$$

Note that all transition probabilities not defined by the above are equal to 0.
The rewards R are defined by:
$R(s, a, s') = \left|(s' - s)^{\frac{1}{3}}\right| \; \forall s \neq s'$,
and $R(s, a, s) = (s+4)^{\frac{-1}{2}}, \forall s \neq 0$.
$R(0, M, 0) = R(0, C, 0) = 0$. Also, the discount factor $\gamma = 0.6$.

We initialize $Q_0(s, a) = 0 \; \forall s \in \{0, 1, 2, 3, 4, 5\}$ and $\forall a \in \{C, M\}$.

1

1/1 point (graded)

We can conclude from this information that 0 is a terminal state.

- (•) True ✔
- ( ) False

Solution:

From the transition probabilities, we can see that no matter which action you take, once you are in state 0, you can never leave.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

2

6.0/6.0 points (graded)

Input the Q-values $Q_1(s, a)$ **correct to 3 decimal places** after one Q-value iteration

$Q_1(0, M) =$ 0 ✔ **Answer:** 0

$Q_1(0, C) =$ 0 ✔ **Answer:** 0

$Q_1(1, M) =$ 1.0 ✔ **Answer:** 1

$Q_1(1, C) =$ 1.0161088135763985 ✔ **Answer:** 1.016

$Q_1(2, M) =$ 1.0 ✔ **Answer:** 1

$Q_1(2, C) =$ 1.00441922206557 ✔ **Answer:** 1.004

$Q_1(3, M) =$ 1.0 ✔ **Answer:** 1

$Q_1(3, C) =$ 0.9953340768291793 ✔ **Answer:** 0.995

$Q_1(4, M) =$ 1.0 ✔ **Answer:** 1

$Q_1(4, C) =$ 0.3535533905932738 ✔ **Answer:** 0.354

$Q_1(5, M) =$ 1.0 ✔ **Answer:** 1

$Q_1(5, C) =$ 0.3333333333333333 ✔ **Answer:** 0.333

Solution:

1. $Q_1(0, M)$: $Q_1(0, M) = 0$ because $R(0, M, 0) = 0$ and $T(0, M, s') = 0 \; \forall s' \neq 0$
2. $Q_1(0, C)$: $Q_1(0, C) = 0$ because $R(0, C, 0) = 0$ and $T(0, C, s') = 0 \; \forall s' \neq 0$
3. $Q_1(1, M)$: $\left|(0-1)^{\frac{1}{3}}\right| = 1$
4. $Q_1(1, C)$: $0.7 * \left|(3-1)^{\frac{1}{3}}\right| + 0.3 * 5^{\frac{-1}{2}} = 0.882 + 0.134 = 1.016$
5. $Q_1(2, M)$: Just as in $Q_1(1, M)$
6. $Q_1(2, C)$: $0.7 * \left|(3-1)^{\frac{1}{3}}\right| + 0.3 * 5^{\frac{-1}{2}} = 0.882 + 0.122 = 1.004$
7. $Q_1(3, M)$: Just as in $Q_1(1, M)$
8. $Q_1(3, C)$: $0.7 * \left|(3-1)^{\frac{1}{3}}\right| + 0.3 * 5^{\frac{-1}{2}} = 0.882 + 0.113 = 0.995$
9. $Q_1(4, M)$: Just as in $Q_1(1, M)$
10. $Q_1(4, C)$: $8^{\frac{-1}{2}} = 0.354$

11. $Q_1(5, M)$: Just as in $Q_1(1, M)$

12. $Q_1(5, C)$: $9^{\frac{-1}{2}} = 0.333$

Submit You have used 2 of 4 attempts

Answers are displayed within the problem

3

3.0/3.0 points (graded)

What are the values $V_1(s)$ corresponding to $Q_1(s, a)$?

$V_1(0) =$ 0 ✔ **Answer:** 0

$V_1(1) =$ 1.0161088135763985 ✔ **Answer:** 1.016

$V_1(2) =$ 1.00441922206557 ✔ **Answer:** 1.004

$V_1(3) =$ 1.0 ✔ **Answer:** 1

$V_1(4) =$ 1.0 ✔ **Answer:** 1

$V_1(5) =$ 1.0 ✔ **Answer:** 1

Solution:

Because: $V_1(s) = \max_a Q_1(s,a)$

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

4

5/5 points (graded)

What are the optimal policies we get from $Q_1(s,a)$?

$\pi^*(1) =$

- (•) C ✔
- ( ) M

$\pi^*(2) =$

- (•) C ✔
- ( ) M

$\pi^*(3) =$

- ( ) C
- (•) M ✔

$\pi^*(4) =$

- ( ) C
- (•) M ✔

$\pi^*(5) =$

- ( ) C

M ✔

Solution:

We pick the policy corresponding to the $V_1(s)$ i.e. $\pi^*(s) = \underset{a}{argmax}\, Q_1(s,a)$

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Homework 6 / 2. Q-Value Iteration

Add a Post

Show all posts ▼ by recent activity ▼

wish the minus sign was a lil' bigger :) 1
spent alot of time on question 2 and before I discovered the minus sign in the second reward function (-1/2) not (1/2).. but hey it is not TOO SMA...

Python solution hints 4
Of course I won't share the whole python script I used but only the most important part of it for those who don't understand how Q and V shoul...

Insights on the policy for the wise

	Insights on the policy for the wise	2
?	More attempts for question 4? Just kidding :D Isn't it kind of useless to give two attempts in question 4? makes it impossible to not get a green check	3
☑	Q(1,C) and Bellman equations	8
💬	hint for excel First I tried to solve it mathematically, but it's a pain to do all these tedious calculations. Then I just made a simple excel and it's really quite simpl...	3
?	Understanding recursion in Q(s,a) Greetings, I'm trying to understand how to calculate recursively. Bellman equations: V(s) = max_a(Q(s,a)) Q(s,a) = ∑_s'(T(s,a,s') * (R(s,a,s') + gamm...	3
?	$T(s,M,s-1)=1$ What is the reward for this? : $R(s,a,s')=\|(s'-s)^{(1/3)}\|$? Is this transition, for example, from state 1 to state 0 ? or 3 to 2?	15
💬	Vectorizing V* and Q*	2
?	[STAFF] Why Gamma for Q2 Do we really need gamma for Q2? The Q_0(s,a) = 0 and so multiplying with gamma is going to be 0. I saved my answers, Can this be checked?	3
?	[STAFF] Q Value Iteration Algorithm I watched course videos twice and I searched for pdf of the lectures. They are not there. I tried to write the Q Value Iteration Algorithm but I feel l...	6
?	Q-Value Iteration 2: why the formulas presented in the lectures don't say anything special about terminating states?	8
💬	About Q(s,a) From lecture Q(s,a) is defined as the expected reward starting at s, taking action a, and acting optimally. I was looking for the state where the rew...	13
?	2. Q-Value Iteration: 2: the values for state 0 not accepted by the grader	5 new 15