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:
$R(s, a, s') = |s' - s|^{\frac{1}{3}} \ \forall s \neq s'$,
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 point possible (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 0 of 1 attempt

Answers are displayed within the problem

2

0.0/6.0 points (graded)

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

$Q_1(0, M) =$ [ ] **Answer:** 0

$Q_1(0, C) =$ [ ] **Answer:** 0

$Q_1(1, M) =$ [ ] **Answer:** 1

$Q_1(1, C) =$ [ ] **Answer:** 1.016

$Q_1(2, M) =$ **Answer:** 1

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

$Q_1(3, M) =$ **Answer:** 1

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

$Q_1(4, M) =$ **Answer:** 1

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

$Q_1(5, M) =$ **Answer:** 1

$Q_1(5, C) =$ **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 0 of 4 attempts

Answers are displayed within the problem

3

0.0/3.0 points (graded)

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

$V_1(0) =$ **Answer:** 0

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

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

$V_1(3) =$ **Answer:** 1

$V_1(4) =$ **Answer:** 1

$V_1(5) =$ **Answer:** 1

Solution:

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

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

4

5 points possible (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 0 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

Thank you for these questions Thank you for these questions i would not have found the algo itself	2
Programmatic Latex display within a loop in Jupyter notebooks	5
Q-value calculation Can someone help me on what I'm doing wrong, based on 1 example Q(4,M); Q(4,M) = T(4,M,...	4
@staff question on V values can i please get more attempts for v values; i think i entered wrong ans in v before calculatin...	2
Terminal state What is the definition of the terminal state? Was it mentioned somewhere in the lecture? I se...	6
[STAFF] Less confusing definition of reward?	5