Course › Unit 5 Reinforcement Learning (2 weeks) › Homework 6 › 2. Q-Value Iteration

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:

$$T\left(s, M, s-1\right) = 1$$

$$T\left(s, C, s+2\right) = 0.7$$

$$T\left(s, C, s\right) = 0.3$$

For state 0:

$$T\left(s, M, s\right) = 1$$

$$T\left(s, C, s\right) = 1$$

For states 4 and 5:

$$T\left(s, M, s-1\right) = 1$$

$$T\left(s, C, s\right) = 1$$

Note that all transition probabilities not defined by the above are equal to 0.

The rewards R are defined by:

$R\left(s, a, s'\right) = \left|\left(s' - s\right)^{\frac{1}{3}}\right| \forall s \neq s',$

and $R\left(s, a, s\right) = \left(s+4\right)^{\frac{-1}{2}}, \forall s \neq 0.$

$R\left(0, M, 0\right) = R\left(0, C, 0\right) = 0$ Also, the discount factor $\gamma = 0.6$.

We initialize $Q_0\left(s, a\right) = 0 \forall s \in \{0,1,2,3,4,5\}$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 ✓ **Answer:** 1

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

$Q_1(2,M) =$ 1 ✓ **Answer:** 1

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

$Q_1(3,M) =$ 1 ✓ **Answer:** 1

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

$Q_1(4,M) =$ 1 ✓ **Answer:** 1

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

$Q_1(5,M) =$ 1 ✓ **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 ✓ **Answer:** 1

$V_1(4) =$ 1 ✓ **Answer:** 1

$V_1(5) =$ 1 ✓ **Answer:** 1

Solution:

Because: $V_1(s) = \underset{a}{max}\, Q_1(s,a)$

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 2 of 2 attempts

Answers are displayed within the problem

Discussion

Show Discussion

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