Course › Unit 5 Reinforcement Learning (2 weeks) › Lecture 18. Reinforcement Learning 2 › 3. Q value iteration by sampling

Audit Access Expires May 11, 2020

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

3. Q value iteration by sampling

Q value iteration by sampling

Video

Download video file

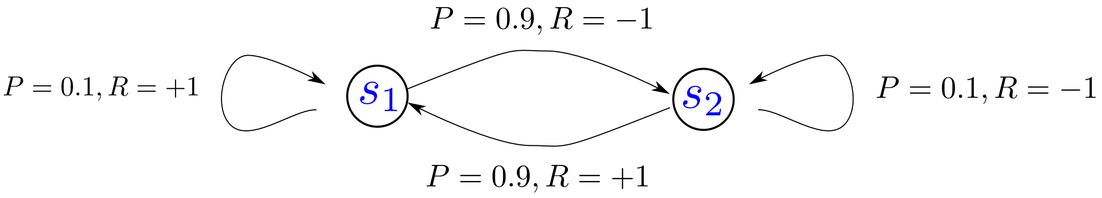
Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Let us consider a toy example which might not be very realistic but which neverthless can help delineate the Q-value iteration for RL using sampling approach.

For this example, assume that there are only two states, s_1, s_2 and only one action possible from each of these states. Let a_{s_1}, a_{s_2} be the actions that could be taken from s_1 and s_2 respectively.



The state transition probabilities are listed below and are also shown in the figure above.

$$T\left(s_1, a_{s_1}, s_1\right) = 0.1$$

$$T\left(s_1, a_{s_1}, s_2\right) = 0.9$$

$$T\left(s_2, a_{s_2}, s_2\right) = 0.1$$

$$T\left(s_2, a_{s_2}, s_1\right) = 0.9$$

The rewards for these actions are given by

$$R\left(s_1, a_{s_1}, s_1\right) = 1$$

$$R\left(s_1, a_{s_1}, s_2\right) = -1$$

$$R\left(s_2, a_{s_2}, s_2\right) = -1$$

$$R\left(s_2, a_{s_2}, s_1\right) = 1$$

Note that we resort to finding optimal Q^* function by sampling for tasks where we don't have access to the exact T, R functions. However, for this toy example we will assume that the Q-value iteration algorithm isn't directly provided with the above specified values of T, R and has to resort to sampling to estimate the Q function.

Let's say that the agent starts out from state s_1 and collects few samples. Each sample can be described by the following tuple $(s, a, s', R(s, a, s'))$ which indicates that the agent received a reward of $R(s, a, s')$ when it reached state s' by taking action a from the state s.

The collected samples are described as follows in the order in which they are presented to the Q-value iteration algorithm.

$$(s_1, a_{s_1}, s_1, +1)$$

$$(s_1, a_{s_1}, s_2, -1)$$

$$(s_2, a_{s_2}, s_1, +1)$$

Let $S_k^{Q(s,a)}$ be used to denote the k^{th} sample of $Q(s,a)$ ($k = i + 1$). Then recall that

$$\hat{Q}_{i+1}(s,a) = \alpha * S_k^{Q(s,a)} + (1 - \alpha) * \hat{Q}_i(s,a)$$

For all of the following problems, assume that the discount factor $\gamma = 0.5$, $\alpha = 0.75$ and that all the Q values are initialized to 0 to start with. That is,

$$\hat{Q}_0(s,a) = 0 \forall s, a$$

Numerical Example

1 point possible (graded)

Enter below the value of $Q(s_1, a_{s_1})$ after the first sample is processed by the Q-value iteration algorithm

Answer: 0.75

Solution:

Let $S_k^{Q(s,a)}$ be used to denote the k^{th} sample of $Q\left(s,a\right)$.

$$\begin{aligned} S_1^{Q(s_1,a_{s_1})} &= R\left(s_1,a_{s_1},s_1\right)+\gamma * \max_{a'} Q\left(s_1,a'\right) \\ S_1^{Q(s_1,a_{s_1})} &= +1+0.5*0=1 \\ Q_1\left(s_1,a_{s_1}\right) &= \alpha * S_1^{Q(s_1,a_{s_1})}+(1-\alpha)*Q_0\left(s_1,a_{s_1}\right) \\ Q_1\left(s_1,a_{s_1}\right) &= .75*1+(1-.75)*0=.75 \end{aligned}$$

Submit

You have used 0 of 3 attempts

Answers are displayed within the problem

Numerical Example - 2

1 point possible (graded)

Enter below the value of $Q\left(s_1,a_{s_1}\right)$ after the second sample is seen by the Q-value iteration algorithm

Answer: -0.5625

Solution:

Let $S_k^{Q(s,a)}$ be used to denote the k^{th} sample of $Q\left(s,a\right)$. Note that from the previous example,

$$Q_1\left(s_1,a_{s_1}\right)=0.75$$

Now we find $S_2^{Q(s_1,a_{s_1})}$:

$$\begin{aligned}
S_2^{Q(s_1,a_{s_1})} &= R\left(s_1, a_{s_1}, s_2\right) + \gamma * \max_{a'} Q\left(s_2, a'\right) \\
S_2^{Q(s_1,a_{s_1})} &= -1 + 0.5 * 0 = -1 \\
Q_2\left(s_1, a_{s_1}\right) &= \alpha * S_2^{Q(s_1,a_{s_1})} + (1-\alpha) * Q_1\left(s_1, a_{s_1}\right) \\
Q_2\left(s_1, a_{s_1}\right) &= 0.75 * -1 + 0.25 * 0.75 = -0.5625
\end{aligned}$$

Submit

You have used 0 of 3 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Lecture 18. Reinforcement Learning 2 / 3. Q value iteration by sampling

Add a Post

Show all posts

by recent activity

Can someone explain the maxQ

Community TA

28

Incorrect formula on board

The last equation on the board in the lecture is incorrect. She has left out a factor of alpha. I....

2

Do STATES increase during RL?

I have a question regarding the many things that are unknown, apart from ***T*** and ***R...

3

Hint for Q1 and Q2

Community TA

3

Heraclitus vs. Markov: Can one ever step in the same river twice?

Is anyone working on a Heraclitus Decision Process (HDP) model?

2

Reinforcement learning resources

Survey paper: https://arxiv.org/abs/cs/9605103 Free textbook: http://incompleteideas.net/bo...

1

? Expression in answers 5

Hi all, should the answers to questions in this vertical be inserted as mathematical expressio...

💬 [Staff] Hand Writing Difficult To Read 3

The hand writing in this particular video is difficult to follow, it is all squished together. Please...

💬 [Staff] Second question bugged grader 2

I introduced an incorrect answer which was graded as correct.

? [Staff] Lecture formula for Q different from exercise? 5

💬 Q value 2

in Max over a' of Q(s_1,a') is the value of Q impacted by the Q initialization to 0?

☑ [STAFF] Not clear how an exponentially weighted average formula leads to the recursive version in the lecture 2