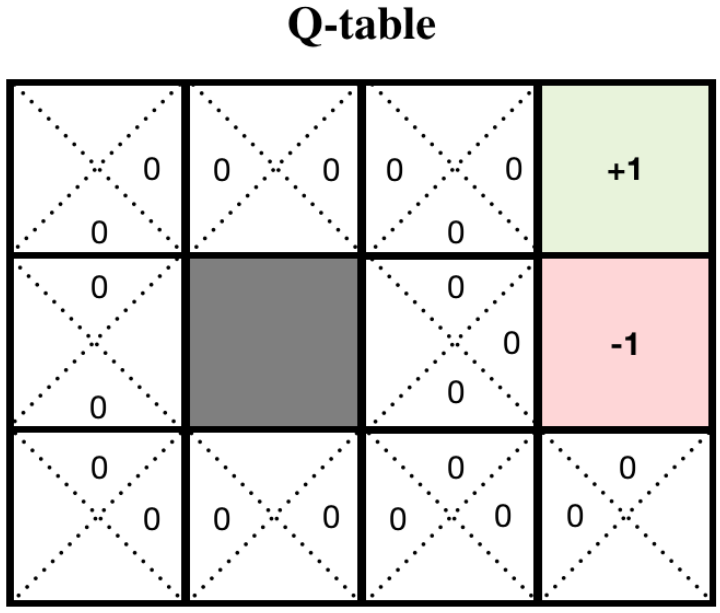
3. Q-Learning

Recall the Q-learning update rule:

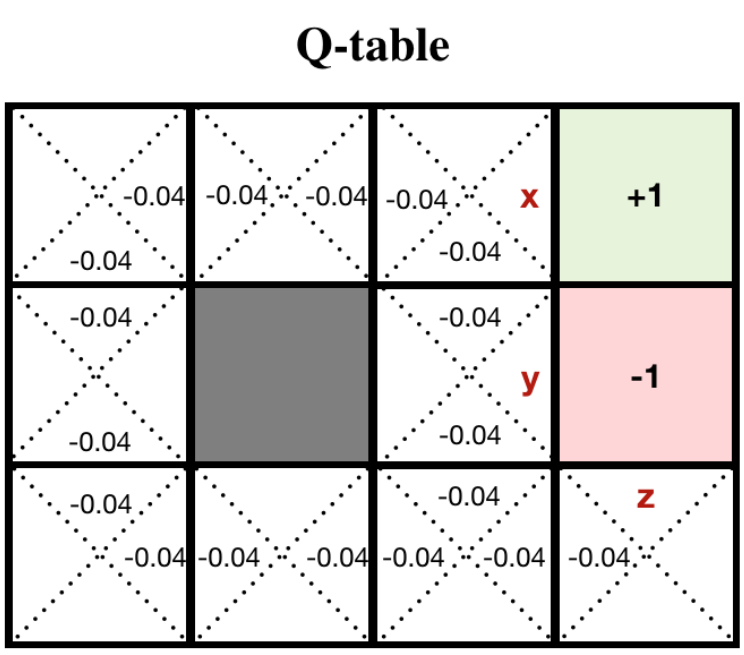
$$Q_{i+1}(s,a) = Q_i(s,a) + \alpha\left[R(s,a,s') + \gamma max_{a'} Q_i(s',a') - Q_i(s,a)\right]$$

let $\alpha = 1$ and $\gamma = 1$ in this problem. In the figure below, at each box, we can go up, down, left and right unless the path is blocked and we initialize the Q value for all the actions in all states as 0. The Q value for the 4 directions are labeled in each box below. Moving into the upper right 2 boxes will result in a reward of $+1$ and -1, and each move will also cost 0.04, or in another word, a reward of -0.04.



1st Iteration

3/3 points (graded)



After 1st iteration, enter the Q value at the position represented by x, y and z below:

$x =$ 0.96 ✔ **Answer:** 0.96

$y =$ -1.04 ✔ **Answer:** -1.04

$z =$ -1.04 ✔ **Answer:** -1.04

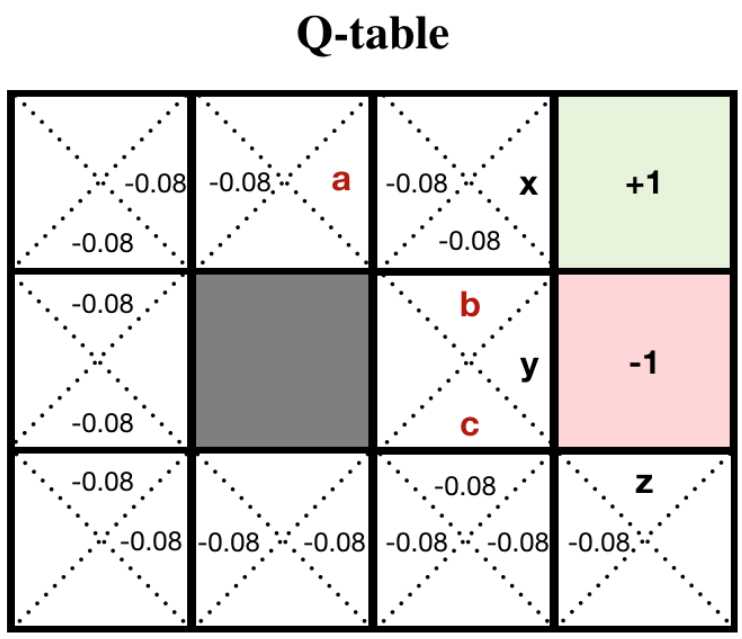
Solution:

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

2nd Iteration

3/3 points (graded)



After 2nd iteration, enter the Q value at the position represented by a, b and c below:

$a =$ -0.08 + 1 ✔ **Answer:** 0.92

$b =$ -0.08 + 1 ✔ **Answer:** 0.92

$c =$ -0.08 ✔ **Answer:** -0.08

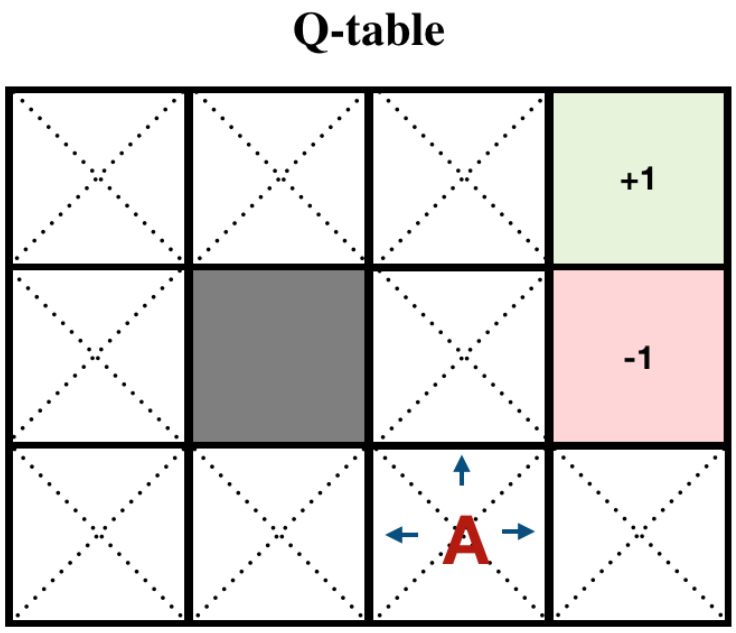
Solution:

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

2nd Iteration

1/1 point (graded)



After convergence, at state A, which action is the optimal?

- ◉ UP ✔
- ○ LEFT
- ○ RIGHT

Solution:

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Epsilon-greedy method 1

0/1 point (graded)
In the ε-greedy method, a larger value of ε would generate experiences that are more consistent with the current Q-value estimates.

- ◉ True ✖
- ○ False ✔

Solution:

In the ε-greedy method, we choose a random action with probability ε and choose an action based on our current estimates with probability 1 - ε. Therefore, it is with smaller ε that we would generate experiences which are more consistent with our current Q-value estimates.

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Epsilon-greedy method 2

1/1 point (graded)
In the ε-greedy method, a value of $\varepsilon = 0.999$ is likely to lead to the desired learning outcome in a highly complex environment.

- ○ True
- ◉ False ✔

Solution:

We would pick a random action virtually every time, and in a highly complex environment, it's highly unlikely that we would properly explore the parts of the space that have high rewards.

在这里，我的理解是，如果环境本身就很复杂，那么本来就很难得到想要的（比如训练机器人）。我们需要足够随机，让他能探索，又得非常好的利用好学到的东西，慢慢往上走。

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Homework 6 / 3. Q-Learning