Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Subject: Reinforcement Learning

AY: 2023 - 24

Experiment 8 (Q Learning Algorithm)

Name: Kresha Shah
SAP ID: 60009220080

AIM:

To implement the Q Learning algorithm in the Grid World environment

THEORY:

Q Learning

Q-learning is a model-free reinforcement learning algorithm to learn the value of an action in a particular state. It does not require a model of the environment (hence "model-free"), and it can handle problems with stochastic transitions and rewards without requiring adaptations. Q-learning is another type of TD method. The 'q' in q-learning stands for quality. Quality in this case represents how useful a given action is in gaining some future reward. The difference between SARSA and Q-learning is that SARSA is an onpolicy model while Q-learning is off-policy.

In the Q-Learning algorithm, the goal is to iteratively learn the optimal Q-value function using the Bellman Optimality Equation. To do so, we store all the Q-values in a table that we will update at each time step using the Q-Learning iteration:

$$q^{new}(s,a) = (1-\alpha)\underbrace{q(s,a)}_{\text{old value}} + \alpha\overbrace{\left(R_{t+1} + \gamma \max_{a'} q(s',a')\right)}^{\text{learned value}}$$

where α is the learning rate, an important hyper parameter that we need to tune since it controls the convergence.

Off-Policy learning:

Off-Policy learning algorithms evaluate and improve a policy that is different from Policy that is used for action selection. In short, [Target Policy != Behaviour Policy]. This helps speed up the convergence i.e. learning can be fast.

ALGORITHM:

Set values for learning rate α、discount rate γ、reward matrix R
Initialize $Q(s,a)$ to zeros
Repeat for each episode,do
    Select state s randomly
    Repeat for each step of episode,do
        Choose a from s using ε-greedy policy or Boltzmann policy
        Take action a obtain reward r from R, and next state s'
        Update $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
        Set $s = s'$
        Until s is the terminal state
    End do
End do

LAB ASSIGNMENT TO DO:

1. Initialize the Grid World environment and implement the Q Learning algorithm
2. Display the initial and final Q-tables
3. Plot the learning curve for different values of alpha (learning rate), gamma (discount factor) and draw your conclusions.

Colab link:
https://colab.research.google.com/drive/1n-Xp8wEdIbEDmwajQuCIUHzYqB2nZWQR?usp=sharing

Output:

```
Gamma=0.1, Alpha=0.1
Initial Q table:
[[[0. 0. 0. 0.]
```

```
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-1.11109125 -1.11109    -1.11109149 -1.11109   ]
  [-1.11090188 -1.1109     -1.11091771 -1.1109    ]
  [-1.10905152 -1.10899999 -1.10912274 -1.10899999]
  [-1.091832   -1.08999999 -1.09116223 -1.09115101]]

 [[-1.11091965 -1.1109     -1.11090574 -1.1109    ]
  [-1.10903691 -1.109      -1.10916325 -1.109     ]
  [-1.09098392 -1.09       -1.09098102 -1.09      ]
  [-0.91326784 -0.9        -0.92210417 -0.90957228]]

 [[-1.10902461 -1.10899999 -1.10916343 -1.10899999]
  [-1.09016091 -1.09       -1.09136817 -1.09      ]
  [-0.90672217 -0.9        -0.9060109  -0.92088976]
  [-0.1         1.          0.          0.        ]]

 [[-1.09098774 -1.09019466 -1.09182378 -1.08999999]
  [-0.90411634 -0.94340929 -0.91748199 -0.9       ]
  [-0.1        -0.43732783 -0.10101     1.        ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.1, Alpha=0.5
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
```

```
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-1.11109815 -1.11109    -1.11109815 -1.11109   ]
  [-1.11096947 -1.1109     -1.11091871 -1.1109    ]
  [-1.1095632  -1.109      -1.10953721 -1.109     ]
  [-1.09419516 -1.09       -1.09904    -1.09419516]]

 [[-1.11100333 -1.1109     -1.11096947 -1.1109    ]
  [-1.10922822 -1.109      -1.10947129 -1.109     ]
  [-1.09745384 -1.09       -1.09985415 -1.09      ]
  [-0.9325     -0.9        -0.9469375  -0.92625   ]]

 [[-1.10947556 -1.109      -1.1095632  -1.109     ]
  [-1.09855357 -1.09       -1.099358   -1.09      ]
  [-0.93753125 -0.9        -0.95570312 -0.9       ]
  [-0.5         1.          0.          0.        ]]

 [[-1.09823976 -1.09419516 -1.09419516 -1.09      ]
  [-0.933125   -0.92625    -0.9469375  -0.9       ]
  [-0.5        -0.5        -0.525       1.        ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.1, Alpha=0.9
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
```

```
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-1.11110118 -1.11109    -1.11110118 -1.11109   ]
  [-1.11105884 -1.1109     -1.11106959 -1.1109    ]
  [-1.1096631  -1.109      -1.11076379 -1.109     ]
  [-1.10349    -1.09       -1.10833307 -1.10349   ]]

 [[-1.11101311 -1.1109     -1.11105884 -1.1109    ]
  [-1.11024873 -1.109      -1.110736   -1.109     ]
  [-1.1049561  -1.09       -1.1082609  -1.09      ]
  [-1.071      -0.9        -0.981      -1.071     ]]

 [[-1.11059764 -1.109097   -1.1096631  -1.109     ]
  [-1.105758   -1.09000045 -1.1080341  -1.09      ]
  [-1.071      -0.909      -0.981      -0.9       ]
  [-0.9         1.          0.          0.        ]]

 [[-1.1043     -1.10349    -1.10349    -1.091511  ]
  [-1.071      -1.071      -1.09449    -0.90000135]
  [-0.9        -0.9        -0.981       0.9999999 ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.5, Alpha=0.1
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
```

```
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-1.9078604  -1.90624976 -1.9078604  -1.90624976]
  [-1.82051034 -1.81249969 -1.82507894 -1.81249966]
  [-1.63194818 -1.62499968 -1.62915766 -1.62499965]
  [-1.28302816 -1.24999996 -1.28745625 -1.28302816]]

 [[-1.82209629 -1.81249967 -1.82051034 -1.81249969]
  [-1.64801242 -1.62499991 -1.62602714 -1.62499991]
  [-1.27102486 -1.24999998 -1.25821782 -1.24999998]
  [-0.55607607 -0.5        -0.56957968 -0.52981622]]

 [[-1.62946388 -1.62499966 -1.63194818 -1.62499968]
  [-1.30564322 -1.24999998 -1.27805251 -1.24999998]
  [-0.59572023 -0.5        -0.50656811 -0.5       ]
  [-0.1         1.          0.          0.        ]]

 [[-1.28410281 -1.28302816 -1.28302816 -1.24999996]
  [-0.55698032 -0.60332541 -0.56937883 -0.5       ]
  [-0.1        -0.1        -0.105       1.        ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.5, Alpha=0.5
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]
```

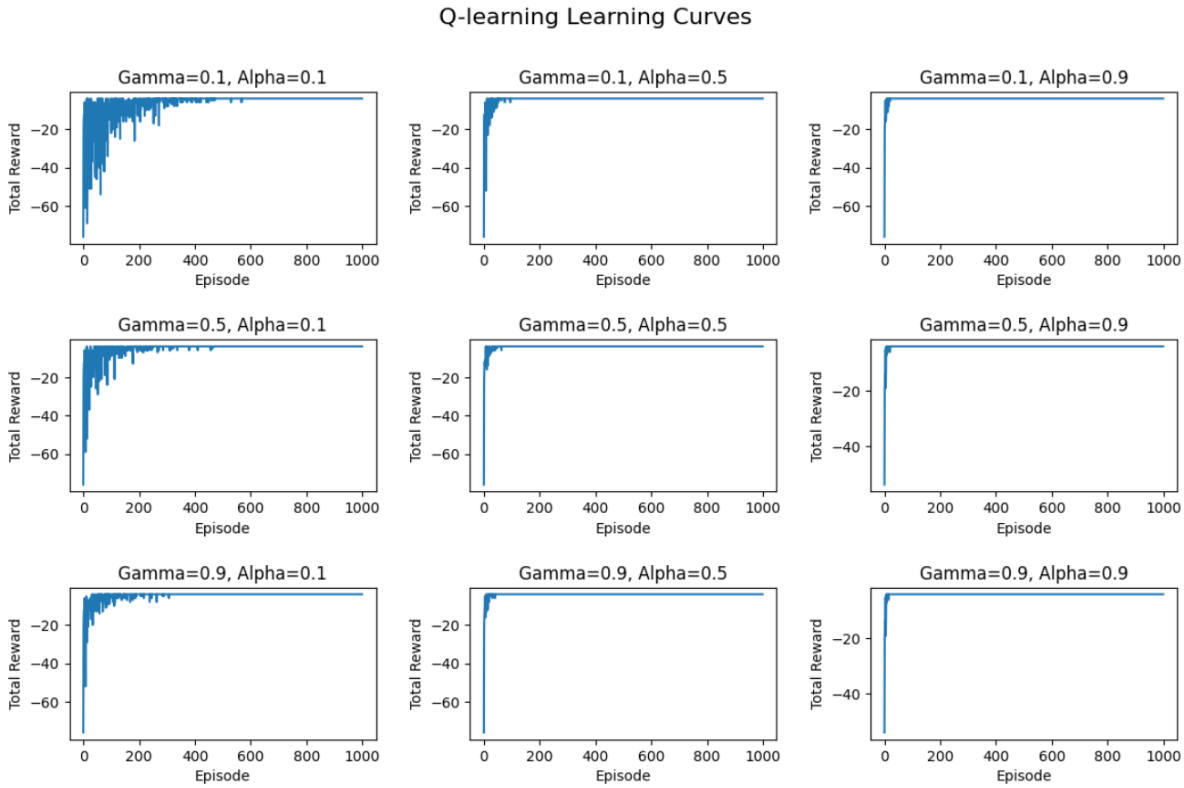
```
Final Q table:
[[[-1.91552973 -1.90625    -1.91552973 -1.90625   ]
  [-1.84983063 -1.8125     -1.88116038 -1.8125    ]
  [-1.64404297 -1.625      -1.69758606 -1.625     ]
  [-1.3671875  -1.25       -1.2734375  -1.3671875 ]]

 [[-1.85051862 -1.8125     -1.84983063 -1.8125    ]
  [-1.7507963  -1.625      -1.69442749 -1.625     ]
  [-1.46630859 -1.25       -1.32519531 -1.25      ]
  [-0.875      -0.5        -0.625      -0.875     ]]

 [[-1.64929199 -1.62503433 -1.64404297 -1.625     ]
  [-1.49954224 -1.25073242 -1.2890625  -1.25      ]
  [-0.875      -0.625      -0.625      -0.5       ]
  [-0.5         1.          0.          0.        ]]

 [[-1.44335938 -1.3671875  -1.3671875  -1.25037766]
  [-0.875      -0.875      -1.03125    -0.50012207]
  [-0.5        -0.5        -0.625       0.99998474]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.5, Alpha=0.9
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-1.94463872 -1.90625    -1.94463872 -1.90625   ]
  [-1.8169875  -1.8125     -1.92342686 -1.8125    ]
  [-1.66725    -1.625      -1.7009775  -1.686375  ]
  [-1.66725    -1.2876525  -1.61775    -1.66725   ]]
```

```
 [[-1.90538372 -1.8125     -1.8169875  -1.8125    ]
  [-1.689525   -1.625      -1.65825    -1.625     ]
  [-1.395      -1.25       -1.305      -1.395     ]
  [-1.395      -0.504945   -1.305      -1.395     ]]

 [[-1.7897625  -1.66725    -1.66725    -1.625     ]
  [-1.395      -1.404      -1.305      -1.25      ]
  [-0.9        -0.9        -1.305      -0.5       ]
  [-0.9         1.          0.          0.        ]]

 [[-1.66725    -1.395      -1.395      -1.395     ]
  [-1.395      -1.395      -1.395      -0.99      ]
  [-0.9        -0.9        -1.305       0.9       ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.9, Alpha=0.1
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-3.50897372 -3.50460994 -3.50897372 -3.50460994]
  [-2.82269467 -2.78289996 -2.87002032 -2.78289996]
  [-1.9836941  -1.98099998 -2.11891398 -1.98099998]
  [-1.13615128 -1.09       -1.13830176 -1.13615128]]

 [[-2.84726965 -2.78289997 -2.82269467 -2.78289997]
  [-2.05383834 -1.981      -2.07213317 -1.981     ]
  [-1.2540706  -1.09       -1.20285549 -1.09      ]
  [-0.41482369 -0.1        -0.43807121 -0.3940399 ]]
```

```
 [[-2.09859509 -1.98162124 -1.9836941  -1.981     ]
  [-1.16921163 -1.09157634 -1.13068775 -1.09      ]
  [-0.31363849 -0.28479326 -0.34246315 -0.1       ]
  [-0.1         1.          0.          0.        ]]

 [[-1.25321691 -1.13615128 -1.13615128 -1.10238372]
  [-0.41544779 -0.3940399  -0.43313487 -0.10556309]
  [-0.1        -0.1        -0.109       0.9991405 ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.9, Alpha=0.5
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-3.6975059  -3.50461    -3.6975059  -3.50461   ]
  [-3.01662704 -2.7829     -3.08827743 -2.7829    ]
  [-2.26219063 -1.981      -2.41153266 -1.99879004]
  [-1.42625    -1.14847778 -1.88848437 -1.42625   ]]

 [[-3.45315477 -2.78318931 -3.01662704 -2.7829    ]
  [-2.07864375 -2.00520166 -2.37022969 -1.981     ]
  [-1.5280625  -1.09       -1.2        -1.19625   ]
  [-0.975      -0.1181797  -0.725      -0.975     ]]

 [[-2.32835227 -2.05228125 -2.26219063 -1.98168301]
  [-1.58375    -1.269375   -1.7418125  -1.09012165]
  [-0.975      -0.525      -0.725      -0.1       ]
  [-0.5         1.          0.          0.        ]]
```

```
 [[-1.4825     -1.42625    -1.42625    -1.32171875]
  [-0.975      -0.975      -1.30125    -0.196875  ]
  [-0.5        -0.5        -0.725       0.984375  ]
  [ 0.          0.          0.          0.        ]]]


Gamma=0.9, Alpha=0.9
Initial Q table:
[[[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]

 [[0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]
  [0. 0. 0. 0.]]]


Final Q table:
[[[-3.75967855 -3.50461    -3.75967855 -3.50461   ]
  [-3.1425039  -2.7829     -3.01312809 -2.7829    ]
  [-2.46429    -1.981      -2.5873119  -2.31849   ]
  [-1.719      -1.728      -2.38239    -1.719     ]]

 [[-3.85804921 -2.7829     -3.1425039  -2.7829    ]
  [-2.530629   -1.9868931  -2.45529    -1.981     ]
  [-1.719      -1.09       -1.629      -1.719     ]
  [-1.719      -0.18981    -1.629      -1.719     ]]

 [[-3.0679749  -2.46429    -2.46429    -1.981     ]
  [-1.719      -1.728      -1.629      -1.09      ]
  [-0.9        -0.9        -1.629      -0.1       ]
  [-0.9         1.          0.          0.        ]]

 [[-2.46429    -1.719      -1.719      -1.719     ]
  [-1.719      -1.719      -1.719      -0.99      ]
  [-0.9        -0.9        -1.629       0.9       ]
   [ 0.          0.          0.          0.        ]]]
```

[5]


Q-learning Learning Curves
Gamma=0.1, Alpha=0.1
Gamma=0.1, Alpha=0.5
Gamma=0.1, Alpha=0.9
Gamma=0.5, Alpha=0.1
Gamma=0.5, Alpha=0.5
Gamma=0.5, Alpha=0.9
Gamma=0.9, Alpha=0.1
Gamma=0.9, Alpha=0.5
Gamma=0.9, Alpha=0.9
Total Reward
Episode