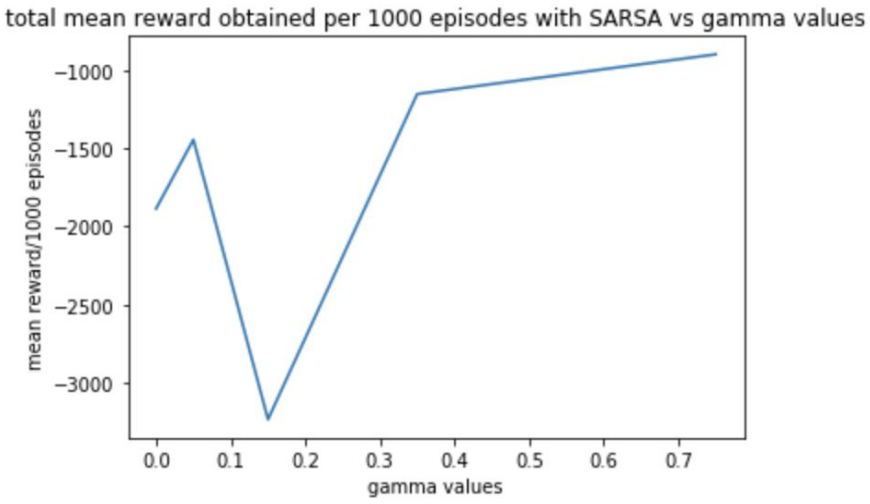
SARSA Gamma Variations

```
State Value Function evaluation for Random Policy with Gamma .99
0 [-22.9, -17.53, -27.23, -25.88]
1 [-23.03, -15.85, -24.46, -26.76]
2 [-19.93, -14.24, -30.66, -23.71]
3 [-18.68, -15.48, -21.96, -24.95]
4 [-16.11, -11.37, -19.71, -21.33]
5 [-17.8, -11.33, -21.15, -19.64]
6 [-12.18, -9.06, -17.19, -15.05]
7 [-13.67, -8.09, -14.79, -15.43]
8 [-12.17, -7.04, -17.51, -12.88]
9 [-13.07, -5.76, -12.62, -13.35]
10 [-6.81, -4.38, -9.62, -15.31]
11 [-6.34, -5.74, -3.22, -9.21]
12 [-21.27, -30.0, -26.71, -24.81]
13 [-22.08, -28.8, -41.84, -31.23]
14 [-20.36, -29.83, -56.86, -33.87]
15 [-17.4, -31.56, -48.63, -33.43]
16 [-15.36, -23.67, -39.44, -37.4]
17 [-15.97, -23.64, -31.81, -26.46]
18 [-20.0, -14.04, -25.23, -28.8]
19 [-11.15, -18.84, -28.29, -21.29]
20 [-12.02, -18.64, -41.64, -21.19]
21 [-9.39, -22.27, -30.18, -19.74]
22 [-10.55, -3.43, -20.58, -17.52]
23 [-8.95, -3.82, -2.17, -6.73]
24 [-22.93, -35.72, -29.4, -28.08]
25 [-28.25, -38.39, -100.0, -42.99]
26 [-28.26, -44.23, -100.0, -63.23]
27 [-31.27, -49.32, -99.98, -47.65]
28 [-27.94, -37.56, -99.92, -57.35]
29 [-28.26, -34.14, -99.67, -39.05]
30 [-23.01, -24.38, -98.62, -29.57]
31 [-23.49, -37.43, -98.62, -30.19]
32 [-20.58, -22.68, -99.03, -32.9]
33 [-12.68, -26.08, -98.62, -27.3]
34 [-15.65, -1.99, -99.94, -25.58]
35 [-4.93, -2.02, -1.0, -3.64]
36 [-24.6, -100.0, -27.87, -36.83]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
```

```
State Value Function evaluation for Random Policy with Gamma .1
0 [-1.11, -1.11, -1.11, -1.11]
1 [-1.11, -1.11, -1.12, -1.11]
2 [-1.11, -1.11, -1.11, -1.11]
3 [-1.11, -1.11, -1.11, -1.11]
4 [-1.11, -1.11, -1.11, -1.11]
5 [-1.11, -1.11, -1.11, -1.11]
6 [-1.11, -1.11, -1.11, -1.11]
7 [-1.11, -1.11, -1.12, -1.11]
8 [-1.11, -1.11, -1.12, -1.11]
9 [-1.11, -1.11, -1.11, -1.11]
10 [-1.11, -1.11, -1.11, -1.11]
11 [-1.11, -1.11, -1.11, -1.11]
12 [-1.11, -1.14, -1.11, -1.11]
13 [-1.11, -1.11, -2.13, -1.11]
14 [-1.11, -1.11, -1.44, -1.15]
15 [-1.11, -1.11, -1.15, -1.11]
16 [-1.11, -1.11, -1.13, -1.11]
17 [-1.11, -1.11, -1.26, -1.11]
18 [-1.11, -1.11, -2.56, -1.11]
19 [-1.11, -1.12, -3.69, -1.11]
20 [-1.11, -1.11, -1.28, -1.11]
21 [-1.11, -1.11, -1.15, -1.12]
22 [-1.11, -1.11, -1.32, -1.11]
23 [-1.11, -1.11, -1.1, -1.11]
24 [-1.11, -1.16, -1.12, -1.11]
25 [-1.11, -3.91, -100.0, -1.16]
26 [-1.11, -1.44, -100.0, -2.17]
27 [-1.11, -1.22, -100.0, -3.06]
28 [-1.11, -1.16, -100.0, -1.88]
29 [-1.11, -1.15, -99.32, -2.46]
30 [-1.11, -1.19, -99.53, -1.98]
31 [-1.11, -1.97, -99.03, -1.28]
32 [-1.11, -2.74, -98.62, -1.69]
33 [-1.11, -1.79, -98.62, -2.4]
34 [-1.11, -1.1, -88.24, -3.34]
35 [-1.11, -1.1, -1.0, -1.23]
36 [-1.11, -100.0, -1.24, -2.57]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
```

```
gamma values: [0.   0.05 0.15 0.35 0.75]
mean reward over 1000 espisodes: [-1885.748 -1445.238 -3232.06  -1153.462  -900.323]
```



Conclusions:

- The discount factor controls how much value is placed on future rewards in the present
- In cliffwalking, a higher discount factor leads to higher mean rewards (a faster solution to the cliffwalking problem since each time step is -1 reward)

SARSA Learning Rate Variations

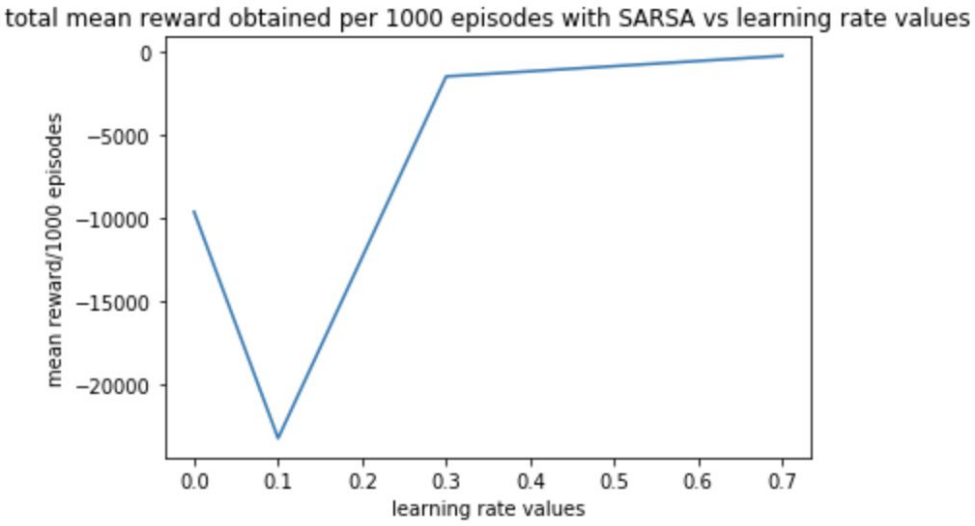
State Value Function evaluation for Random Policy with Learning Rate .00001

```
 0 [-1.0, -1.0, -1.0, -1.0]
 1 [-1.0, -1.0, -1.0, -1.0]
 2 [-1.0, -1.0, -1.0, -1.0]
 3 [-1.0, -1.0, -1.0, -1.0]
 4 [-1.0, -1.0, -1.0, -1.0]
 5 [-1.0, -1.0, -1.0, -1.0]
 6 [-1.0, -1.0, -1.0, -1.0]
 7 [-1.0, -1.0, -1.0, -1.0]
 8 [-1.0, -1.0, -1.0, -1.0]
 9 [-1.0, -1.0, -1.0, -1.0]
10 [-1.0, -1.0, -1.0, -1.0]
11 [-1.0, -1.0, -1.0, -1.0]
12 [-1.0, -1.0, -1.0, -1.0]
13 [-1.0, -1.0, -1.0, -1.0]
14 [-1.0, -1.0, -1.0, -1.0]
15 [-1.0, -1.0, -1.0, -1.0]
16 [-1.0, -1.0, -1.0, -1.0]
17 [-1.0, -1.0, -1.0, -1.0]
18 [-1.0, -1.0, -1.0, -1.0]
19 [-1.0, -1.0, -1.0, -1.0]
20 [-1.0, -1.0, -1.0, -1.0]
21 [-1.0, -1.0, -1.0, -1.0]
22 [-1.0, -1.0, -1.0, -1.0]
23 [-1.0, -1.0, -1.0, -1.0]
24 [-1.0, -1.0, -1.0, -1.0]
25 [-1.0, -1.0, -100.0, -1.0]
26 [-1.0, -1.0, -100.0, -1.0]
27 [-1.0, -1.0, -100.0, -1.0]
28 [-1.0, -1.0, -100.0, -1.0]
29 [-1.0, -1.0, -100.0, -1.0]
30 [-1.0, -1.0, -100.0, -1.0]
31 [-1.0, -1.0, -100.0, -1.0]
32 [-1.0, -1.0, -100.0, -1.0]
33 [-1.0, -1.0, -100.0, -1.0]
34 [-1.0, -1.0, -100.0, -1.0]
35 [-1.0, -1.0, -1.0, -1.0]
36 [-1.0, -100.0, -1.0, -1.0]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
```

State Value Function evaluation for Random Policy with Learning Rate .00001 and 10,000 episodes

```
0 [-0.14, -0.14, -0.14, -0.14]
1 [-0.13, -0.13, -0.13, -0.13]
2 [-0.12, -0.12, -0.12, -0.12]
3 [-0.1, -0.1, -0.1, -0.1]
4 [-0.09, -0.09, -0.09, -0.09]
5 [-0.08, -0.08, -0.08, -0.08]
6 [-0.07, -0.07, -0.07, -0.07]
7 [-0.06, -0.06, -0.06, -0.06]
8 [-0.05, -0.05, -0.05, -0.05]
9 [-0.04, -0.04, -0.04, -0.04]
10 [-0.03, -0.03, -0.03, -0.03]
11 [-0.03, -0.03, -0.03, -0.03]
12 [-0.16, -0.16, -0.16, -0.16]
13 [-0.14, -0.14, -0.14, -0.14]
14 [-0.12, -0.12, -0.12, -0.12]
15 [-0.1, -0.1, -0.1, -0.1]
16 [-0.09, -0.09, -0.09, -0.09]
17 [-0.08, -0.08, -0.08, -0.08]
18 [-0.06, -0.06, -0.06, -0.06]
19 [-0.05, -0.05, -0.05, -0.05]
20 [-0.05, -0.05, -0.05, -0.05]
21 [-0.04, -0.04, -0.04, -0.04]
22 [-0.03, -0.03, -0.03, -0.03]
23 [-0.03, -0.03, -0.03, -0.03]
24 [-0.18, -0.18, -0.18, -0.18]
25 [-0.14, -0.14, -1.22, -0.14]
26 [-0.12, -0.12, -0.96, -0.12]
27 [-0.1, -0.1, -0.83, -0.1]
28 [-0.08, -0.08, -0.67, -0.08]
29 [-0.07, -0.07, -0.54, -0.07]
30 [-0.06, -0.06, -0.45, -0.06]
31 [-0.05, -0.05, -0.38, -0.05]
32 [-0.04, -0.04, -0.35, -0.04]
33 [-0.04, -0.04, -0.28, -0.04]
34 [-0.03, -0.03, -0.23, -0.03]
35 [-0.02, -0.02, -0.02, -0.02]
36 [-0.24, -2.19, -0.24, -0.24]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
```

```
learning rate values: [0.  0.1 0.3 0.7]
mean reward over 1000 espisodes: [ -9637.024 -23242.684  -1501.443   -270.078]
```



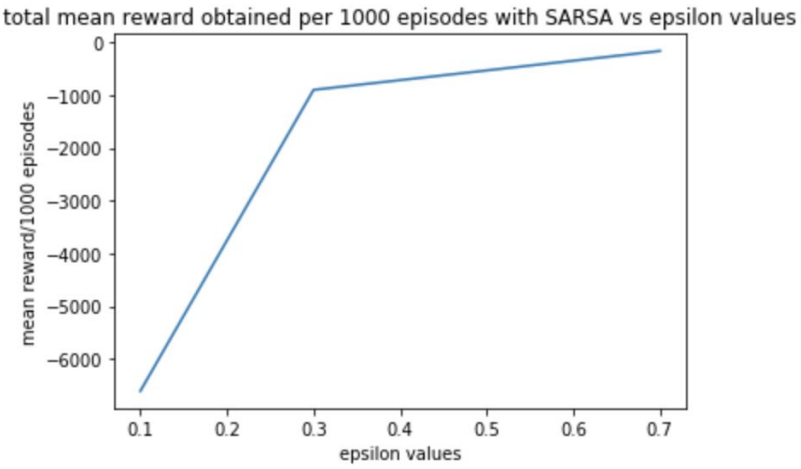
Conclusions:

- Learning rate is the size of the step we adjust values with towards the true mean in each episode
- A large learning rate (close to 1 or over) leads makes changes that are too large in magnitude to converge on an optimal solution

- A learning rate that is too small does not lead to enough change (the agent is not learning enough)
- A tiny learning rate (.001) requires more episodes to converge

SARSA Epsilon Variation

```
epsilon values [0.1 0.3 0.7]
mean reward over 1000 espisodes: [-6613.387  -896.87   -158.657]
```



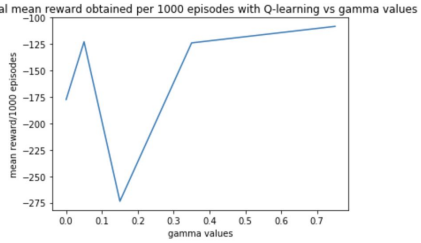
Conclusions:

- Epislon controls the likelihood of choosing an action following the greedy policy
- A high epsilon makes the agent converge on a greedy (optimal) policy
- A low epsilon leads to less reward due to more time exploring

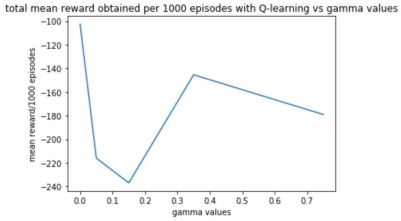
Q-Learning Gamma Variations

```
Q Values for Q-learning after 1000 episodes .9 discount factor
0 [-7.06, -7.06, -7.08, -7.08]
1 [-6.9, -6.89, -6.9, -6.89]
2 [-6.68, -6.68, -6.68, -6.68]
3 [-6.43, -6.41, -6.42, -6.43]
4 [-6.13, -6.12, -6.19, -6.11]
5 [-5.81, -5.8, -5.81, -5.8]
6 [-5.44, -5.43, -5.43, -5.44]
7 [-5.03, -5.02, -5.02, -5.03]
8 [-4.57, -4.56, -4.56, -4.58]
9 [-4.06, -4.05, -4.05, -4.08]
10 [-3.51, -3.48, -3.49, -3.54]
11 [-2.88, -2.93, -2.83, -2.98]
12 [-7.21, -7.22, -9.04, -7.23]
13 [-6.99, -6.98, -10.07, -6.99]
14 [-6.77, -6.87, -7.57, -6.77]
15 [-6.38, -6.38, -10.57, -6.42]
16 [-6.06, -6.06, -8.51, -6.07]
17 [-5.75, -5.73, -7.62, -5.75]
18 [-5.37, -5.37, -8.33, -5.43]
19 [-4.89, -4.87, -6.5, -4.91]
20 [-4.33, -4.3, -5.6, -4.36]
21 [-3.77, -3.64, -4.5, -3.88]
22 [-3.25, -2.81, -3.88, -3.2]
23 [-2.71, -2.43, -1.92, -2.54]
24 [-7.46, -9.9, -24.22, -8.56]
25 [-6.26, -7.19, -83.92, -7.11]
26 [-6.01, -9.54, -51.86, -11.3]
27 [-4.65, -6.07, -69.51, -4.65]
28 [-4.69, -6.31, -58.66, -4.65]
29 [-4.74, -4.49, -56.06, -7.32]
30 [-4.07, -5.45, -54.7, -5.41]
31 [-3.19, -3.53, -53.3, -4.18]
32 [-3.19, -3.17, -43.94, -3.39]
33 [-2.73, -3.3, -42.2, -3.07]
34 [-1.87, -1.85, -38.57, -2.26]
35 [-2.23, -1.71, -1.0, -2.75]
36 [-8.0, -99.94, -16.93, -18.88]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
```

```
gamma values: [0.   0.05 0.15 0.35 0.75]
mean reward over 1000 espisodes: [-177.582 -123.037 -273.553 -124.12  -108.527]
```




```
gamma values: [0.   0.05 0.15 0.35 0.75]
mean reward over 1000 espisodes: [-102.458 -215.965 -236.992 -145.41  -179.042]
```



Conclusions:

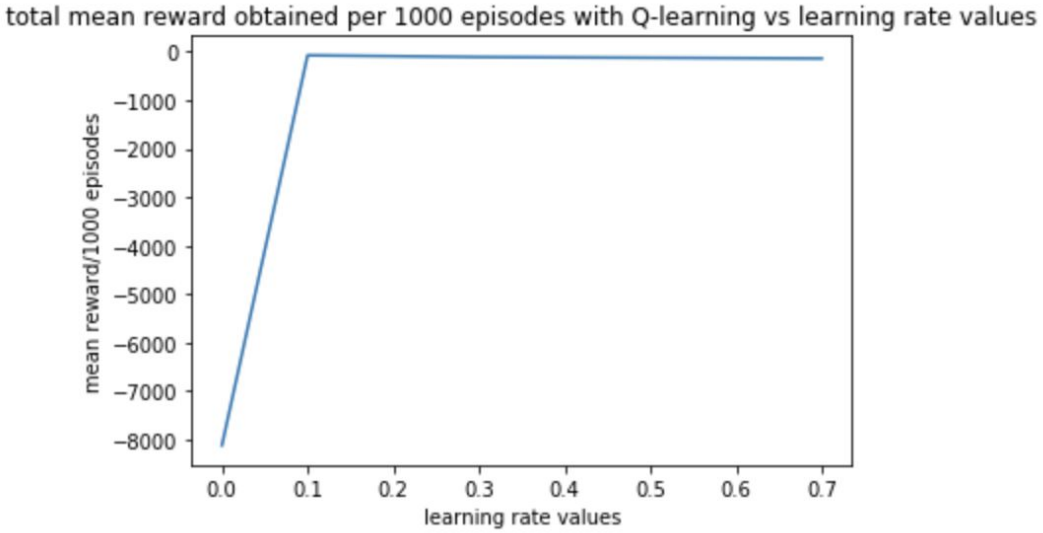
- Q learning in general leads to higher rewards
- Q learning in general converges faster on an optimal (greedy policy)
- A higher discount factor leads to generally higher rewards (although in the experiments there could be a lot of variance between trials)

Q-learning Learning Rate variations

```
Q values with Q learning after 1000 epsisodes with .01 learning rate
0 [-2.84, -2.85, -2.84, -2.84]
1 [-2.73, -2.73, -2.73, -2.72]
2 [-2.56, -2.56, -2.57, -2.56]
3 [-2.37, -2.38, -2.38, -2.38]
4 [-2.18, -2.19, -2.19, -2.18]
5 [-2.0, -2.0, -2.0, -1.99]
6 [-1.81, -1.81, -1.82, -1.81]
7 [-1.63, -1.63, -1.64, -1.64]
8 [-1.46, -1.46, -1.45, -1.47]
9 [-1.29, -1.29, -1.29, -1.29]
10 [-1.15, -1.15, -1.15, -1.16]
11 [-1.05, -1.06, -1.05, -1.06]
12 [-2.96, -2.96, -2.96, -2.96]
13 [-2.78, -2.79, -2.78, -2.79]
14 [-2.58, -2.58, -2.59, -2.58]
15 [-2.38, -2.38, -2.37, -2.38]
16 [-2.18, -2.17, -2.17, -2.17]
17 [-1.98, -1.98, -1.98, -1.98]
18 [-1.79, -1.79, -1.8, -1.8]
19 [-1.61, -1.61, -1.61, -1.61]
20 [-1.43, -1.43, -1.43, -1.43]
21 [-1.25, -1.25, -1.25, -1.27]
22 [-1.09, -1.09, -1.09, -1.09]
23 [-0.95, -0.95, -0.95, -0.95]
24 [-3.22, -3.22, -3.23, -3.22]
25 [-2.86, -2.86, -77.18, -2.87]
26 [-2.59, -2.59, -64.48, -2.59]
27 [-2.36, -2.36, -51.99, -2.36]
28 [-2.14, -2.15, -45.83, -2.15]
29 [-1.94, -1.95, -40.1, -1.95]
30 [-1.76, -1.76, -38.89, -1.76]
31 [-1.56, -1.57, -29.66, -1.57]
32 [-1.38, -1.39, -28.23, -1.38]
33 [-1.19, -1.19, -23.0, -1.2]
34 [-0.99, -0.98, -15.71, -0.98]
35 [-0.68, -0.68, -0.68, -0.68]
36 [-3.66, -94.69, -3.66, -3.66]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
40 [0, 0, 0, 0]
```

```
Q values with Q learning after 10000 epsisodes with .01 learning rate
0 [-7.47, -7.47, -7.47, -7.47]
1 [-7.31, -7.31, -7.32, -7.32]
2 [-7.11, -7.11, -7.11, -7.12]
3 [-6.88, -6.88, -6.88, -6.88]
4 [-6.62, -6.62, -6.62, -6.62]
5 [-6.33, -6.33, -6.33, -6.33]
6 [-6.02, -6.02, -6.02, -6.02]
7 [-5.68, -5.68, -5.68, -5.68]
8 [-5.33, -5.32, -5.32, -5.32]
9 [-4.96, -4.95, -4.95, -4.95]
10 [-4.57, -4.57, -4.57, -4.57]
11 [-4.24, -4.24, -4.24, -4.24]
12 [-7.62, -7.61, -7.62, -7.62]
13 [-7.42, -7.42, -7.42, -7.42]
14 [-7.19, -7.19, -7.19, -7.19]
15 [-6.94, -6.93, -6.94, -6.94]
16 [-6.65, -6.65, -6.66, -6.66]
17 [-6.34, -6.34, -6.35, -6.35]
18 [-6.01, -6.0, -6.0, -6.0]
19 [-5.64, -5.63, -5.64, -5.64]
20 [-5.24, -5.24, -5.24, -5.24]
21 [-4.81, -4.81, -4.81, -4.81]
22 [-4.35, -4.35, -4.35, -4.35]
23 [-3.89, -3.88, -3.88, -3.89]
24 [-7.81, -7.81, -7.91, -7.82]
25 [-7.58, -7.58, -96.07, -7.59]
26 [-7.32, -7.32, -95.47, -7.4]
27 [-7.03, -7.03, -90.76, -7.05]
28 [-6.72, -6.72, -89.89, -6.75]
29 [-6.37, -6.37, -85.19, -6.4]
30 [-5.99, -5.99, -85.33, -6.03]
31 [-5.57, -5.56, -81.52, -5.57]
32 [-5.12, -5.09, -83.95, -5.12]
33 [-4.58, -4.55, -84.58, -4.62]
34 [-3.97, -3.96, -82.78, -3.97]
35 [-3.31, -2.86, -1.0, -3.47]
36 [-8.03, -99.78, -8.03, -8.03]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
```

```
learning rate values: [0.  0.1 0.3 0.7]
mean reward over 1000 espisodes: [-8116.213   -71.102  -106.465  -136.774]
```



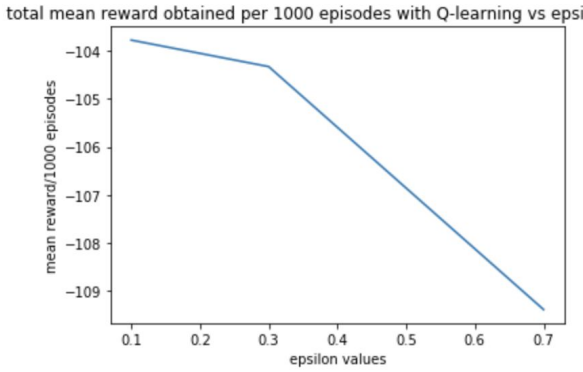
Conclusions:

- In Q learning, the learning rate had slightly less influence on the convergence towards higher rewards
- However similar effects in convergence as in SARSA are noted, a smaller learning rate required more trials to converge

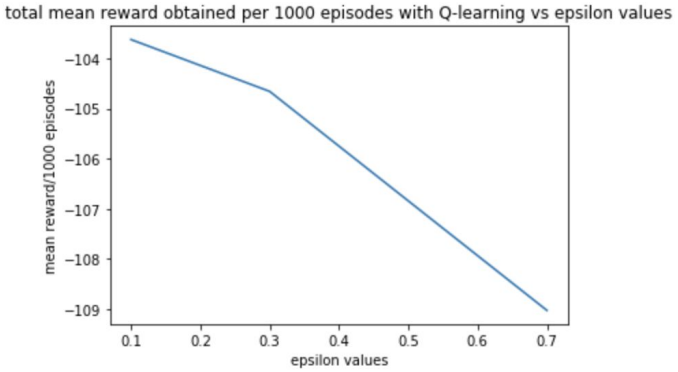
Q-learning epsilon variation

```
Q values with Q learning after 1000 epsisodes with .1 epsilon
0 [-4.72, -4.72, -4.73, -4.73]
1 [-4.62, -4.62, -4.61, -4.62]
2 [-4.45, -4.46, -4.46, -4.45]
3 [-4.26, -4.26, -4.26, -4.26]
4 [-4.05, -4.04, -4.05, -4.05]
5 [-3.81, -3.8, -3.81, -3.81]
6 [-3.55, -3.55, -3.55, -3.55]
7 [-3.28, -3.28, -3.28, -3.28]
8 [-2.98, -2.98, -2.97, -2.98]
9 [-2.66, -2.66, -2.65, -2.66]
10 [-2.34, -2.34, -2.34, -2.34]
11 [-2.09, -2.09, -2.09, -2.09]
12 [-4.82, -4.83, -4.83, -4.83]
13 [-4.67, -4.67, -4.67, -4.67]
14 [-4.48, -4.48, -4.48, -4.48]
15 [-4.26, -4.26, -4.27, -4.28]
16 [-4.04, -4.04, -4.04, -4.04]
17 [-3.8, -3.8, -3.8, -3.8]
18 [-3.54, -3.54, -3.54, -3.54]
19 [-3.26, -3.25, -3.25, -3.25]
20 [-2.93, -2.93, -2.94, -2.94]
21 [-2.58, -2.58, -2.58, -2.58]
22 [-2.19, -2.18, -2.18, -2.19]
23 [-1.78, -1.78, -1.77, -1.78]
24 [-5.04, -5.04, -5.05, -5.04]
25 [-4.75, -4.75, -45.83, -4.75]
26 [-4.5, -4.5, -41.88, -4.5]
27 [-4.27, -4.27, -34.43, -4.27]
28 [-4.04, -4.04, -22.22, -4.03]
29 [-3.78, -3.78, -35.74, -3.78]
30 [-3.52, -3.52, -28.23, -3.51]
31 [-3.22, -3.22, -24.53, -3.22]
32 [-2.88, -2.88, -24.53, -2.88]
33 [-2.46, -2.45, -22.22, -2.46]
34 [-1.87, -1.87, -18.21, -1.87]
35 [-1.06, -1.02, -1.0, -1.16]
36 [-5.38, -57.87, -5.37, -5.38]
37 [0, 0, 0, 0]
38 [0, 0, 0, 0]
39 [0, 0, 0, 0]
```

```
epsilon values: [0.1 0.3 0.7]
mean reward over 1000 espisodes: [-103.774 -104.326 -109.386]
```




```
epsilon values: [0.1 0.3 0.7]
mean reward over 1000 espisodes: [-103.627 -104.66  -109.026]
```



Conclusions:

In cliff walking and Q learning there were slightly lower rewards associated with a high epsilon (the likelihood of choosing a random action)